Science Library

OCT 11 1948

Vol. 45, No. 5 September, 1948

Psychological Bulletin

EDITED BY
LYLE H. LANIER, NEW YORK UNIVERSITY

179

WITH THE CO-OPERATION OF
S. H. BRITT, McCANN-ERICKSON, INC., NEW YORK; D. A. GRANT, UNIVERSITY OF WISCONSIN; W. T. HERON, UNIVERSITY OF MINNESOTA; W. A. HUNT, NORTHWESTERN UNIVERSITY; D. G. MARQUIS, UNIVERSITY OF MICHIGAN; A. W. MELTON, OHIO STATE UNIVERSITY; J. T. METCALF, UNIVERSITY OF VERMONT.

CONTENTS

PUBLISHED BIMONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1515 Massachusetts Ave., N.W., Washington 5, D.C.

Subscription price, $7 to per year, single issue, $1.25.

Entered as second class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage provided for in Section 538, act of February 28, 1925, authorized August 6, 1918.

PUBLICATIONS OF

The American Psychological Association, Inc.

AMERICAN PSYCHOLOGIST

Editor: DAEL WOLFLE, *American Psychological Association*

Contains all official papers of the Association and articles concerning psychology as a profession; monthly.

Subscription: $7.00 (Foreign $7.50). Single copies, $.65.

JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY

Editor: GORDON W. ALLPORT, *Harvard University*

Contains original contributions in the field of abnormal and social psychology, reviews, and case reports; quarterly.

Subscription: $5.00 (Foreign $5.25). Single copies, $1.50.

JOURNAL OF APPLIED PSYCHOLOGY

Editor: DONALD G. PATERSON, *University of Minnesota*

Contains material covering applications of psychology to business, industry, education, etc.; bi-monthly.

Subscription: $6.00 (Foreign $6.50). Single copies, $1.25.

JOURNAL OF COMPARATIVE AND PHYSIOLOGICAL PSYCHOLOGY

Editor: CALVIN P. STONE, *Stanford University*

Contains original contributions in the field of comparative and physiological psychology; bi-monthly.

Subscription: $7.00 (Foreign $7.50). Single copies, $1.25.

JOURNAL OF CONSULTING PSYCHOLOGY

Editor: LAURANCE F. SHAFFER, *Teachers College, Columbia University*

Contains articles in the field of clinical and consulting psychology, counseling and guidance; bi-monthly.

Subscription: $5.00 (Foreign $5.50). Single copies, $1.00.

JOURNAL OF EXPERIMENTAL PSYCHOLOGY

Editor: FRANCIS W. IRWIN, *University of [illegible]*

Contains original contributions of an experimental character; bi-monthly.

Subscription: $7.00 (Foreign $7.25). Single copies, $1.25.

PSYCHOLOGICAL ABSTRACTS

Editor: C. M. LOUTTIT, *University of Illinois, Galesburg, Illinois*

Contains noncritical abstracts of the world's literature in psychology and related subjects; monthly.

Subscription: $7.00 (Foreign $7.25). Single copies, $.75.

PSYCHOLOGICAL BULLETIN

Editor: LYLE H. LANIER, *New York University*

Contains critical reviews of books and articles and critical and analytical summaries of psychological fields or subject matter; bi-monthly.

Subscription: $7.00 (Foreign $7.25). Single copies, $1.25.

PSYCHOLOGICAL MONOGRAPHS: GENERAL AND APPLIED

Editor: HERBERT S. CONRAD, *College Entrance Examination Board*

Contains longer researches and laboratory studies which appear as units; published at irregular intervals at a cost to author of about $2.50 a page; author receives 150 copies gratis.

Subscription: $6.00 per volume of about 550 pages (Foreign $6.50). Single copies, price varies according to size.

PSYCHOLOGICAL REVIEW

Editor: CARROLL C. PRATT, *Princeton University*

Contains original contributions of a theoretical nature; bi-monthly.

Subscription: $5.50 (Foreign $5.75). Single copies, $.75.

Subscriptions are payable in advance and are terminated [illegible]

Make checks payable to the American Psychological Association, Inc.

Subscriptions, orders, and other business communications should be sent to

AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

1515 MASSACHUSETTS AVENUE N.W., WASHINGTON 5, D.C.

GEORGE BANTA PUBLISHING COMPANY, MENASHA, WISCONSIN

Vol. 45, No. 5 September, 1948

Psychological Bulletin

THE VALIDITY OF PERSONALITY INVENTORIES IN MILITARY PRACTICE[1]

ALBERT ELLIS
Northern New Jersey Mental Hygiene Clinic
AND
HERBERT S. CONRAD
Educational Testing Service

INTRODUCTION

One of the most significant military contributions to psychology has been the development of personality inventories. Such inventories have proved helpful for the neuropsychiatric screening of adults both prior to induction into military service, and after induction. In contrast, the use of personality inventories in civilian practice for the past thirty years has generally yielded disappointing results.[2]

The question thus arises: What are the reasons for the superior showing of the personality inventories in military practice? To throw some light on this matter, the present authors undertook to review the available papers on military validation of personality questionnaires.[3] It was considered that such a review should (1) summarize the findings

[1] The writers wish to express their gratitude to several colleagues who have critically read this paper and offered valuable suggestions: Dr. Walter C. Shipley, of Wheaton College; Drs. Silvan S. Tomkins, Glenn V. Ramsey, and Norman Frederiksen, of Princeton University; and Dr. William B. Schrader, of the Educational Testing Service, Princeton, N. J. Any errors that may remain are, of course, our own responsibility.

[2] See the reviews by Ellis (12, 13), Gilliland (16), Maller (32), Roback (42), Rosenzweig (43), Thorpe (57), Traxler (58), Vernon (60), and Wiley and Trimble (67). Kornhauser (25), questioning 67 well-known psychologists as to how satisfactory they considered personality inventories for individual classification, found that only 1.5% of his respondents upheld the inventories as highly satisfactory; 13.5% considered them moderately satisfactory; while the remainder regarded the inventories as "doubtfully satisfactory," "rather unsatisfactory," or "highly unsatisfactory."

[3] Papers on the type of questionnaire known as the Biographical Data Inventory (such as that used by Fiske (14)), are not included in the present review. The biographical inventory generally concentrates on historical data rather than current adjustment problems, and is usually quite different from the conventional personality or adjustment inventory.

from the military applications of personality inventories; (2) describe both the legitimate and spurious sources of such military superiority as is observed; and (3) point out the lessons for civilian work in this field. The present review is limited to these three purposes. No attempt will be made, for example, to evaluate the merits of one inventory as against another.[4]

Types of Validation

There were, in general, two important methods of estimating the value of personality inventories in military practice: (1) by use of a psychiatric criterion—i.e., comparing the inventory scores of "normal" groups of enlisted men or officers with the scores of others prognosed or diagnosed as neuropsychiatrically unfit; and (2) by use of performance-measures—i.e., comparing the inventory scores of those who were successful in training or combat, with the scores of those who were failures. The results from these two methods of validation will be considered separately.

Military Validation Making Use of Psychiatric Criteria

A number of military studies were found in which psychiatric criteria (either prognoses or diagnoses of unfitness) were employed in the validation of personality inventories. The basic data of these reports are summarized in Table I. Examination of the last column of Table I will show that in only a handful of instances were definitely unfavorable or negative results obtained, while in the overwhelming majority of studies the instrument in question proved to have some value for screening or diagnostic purposes. The near-unanimity of favorable results is impressive—and all the more so when one remembers that psychiatric prognoses themselves are by no means uniformly reliable and valid.[5] How-

[4] With regard to this problem, Wexler, reporting on studies in which different inventories were employed in identical samples, concludes: "In no instance has it appeared with clarity and certainty that a particular instrument, a particular content, or a particular format is decidedly superior to any other instrument, content, or format" (65, p. 169).

[5] Two pertinent studies may be cited. (1) Weinstock and Watson (64) reported on 121 Naval recruits who were allowed to remain in service ("on trial") despite an adverse prognosis based on clinical judgment. Of the 121 recruits, only 44 (or 36%) were discharged for neuropsychiatric reasons during recruit training. This suggests a rather high "false positive" ratio—which casts doubt on the validity of the clinical prognosis as a criterion. (2) Varney and Stone (59), attacking this problem from another angle, report a study involving 813 Maritime Service trainees disenrolled for neuropsychiatric causes. Of the 813 disenrollees, 247 (or 30%) had passed through a series of psychiatric screening

ever, some of the validity figures are so high—.75 for the Bernreuter Personality Inventory ($N = 200$), .78 for the Psychoneurotic Inventory ($N = 600$), and .80 for the Psychosomatic Inventory ($N = 200$)[6]—that they suggest the operation of unusual experimental or statistical factors. In the sections below, reasons (both spurious and legitimate) for the superior validity of personality inventories in military practice will be considered. These reasons are not offered as original with the writers, nor does the listing of spurious reasons imply that the original investigators failed to understand the limitations of their findings. It has seemed worthwhile, however, to assemble the various reasons or considerations in one place, for convenient reference in the evaluation of research in this field.

1. *Criterion contamination.* In several of the studies there was either some possibility of, or direct evidence of, criterion contamination. That is to say, the psychiatrists, psychologists, or officers who made the ratings which served as criteria of the respondents' maladjustment or failure knew the respondents' inventory scores, and probably allowed these scores to influence their judgments. To the extent that this occurs, the obtained validity coefficients are, of course, too high.

Thus, in Coville's study (**10**) 77% of the psychiatric disenrollees serving as a criterion group for the Maritime Service Inventory were selected by means of an "admission examination." In the main, this admission examination consisted, first, of the MSI; and second (for those screened by the MSI) of a neuropsychiatric interview. MSI scores were freely available for reference during the interview. There would appear to be appreciable likelihood of criterion-contamination here, even though referral for interview (based mainly on MSI score) was at the rate of 30–40% of the examinees, whereas actual rejection for "psychiatric and neurological disorders" was less than 1% (**23**, pp. 102–103, 304).—In Gough's study (**17**) mention is made of the possibility of criterion contamination in some of the psychiatric diagnoses.—In the investigation by Miles and others[7] (**34**), the psychiatrists making the original diagnoses were aware of the subjects' Self-Descriptive Inventory scores; and the drill instructors who finally judged some of the subjects as unfit were aware of the psychiatrists' ratings. Since the drill instructors' judgments served as the criterion, it is evident that the possibility of criterion-contamination here may have been significant.

---

processes without rejection. Here the proportion of "false negatives" is perhaps disturbingly high. Unfortunately, there are only relatively few studies which, like the ones cited, have made use of a follow-up record of actual maladjustment (as distinguished from a prognosis alone). While the practical usefulness of the psychiatric prognosis is well substantiated, the use of the prognosis alone, as a criterion in scientific research, clearly leaves much to be desired.

[6] Summaries of these studies are given in the last three entries of Table I.

[7] See the 24th entry in Table II.

TABLE I

MILITARY VALIDITY STUDIES OF PERSONALITY INVENTORIES, MAKING USE OF PSYCHIATRIC CRITERIA

Source	*Group Tested*	*Criterion*	*Result*
		Personal Inventory	
Bobbitt & Newman (7)	99 Coast Guard trainees	Patterns of hospital complaints of men obtaining "normal" and "abnormal" PI scores	"The patterns of hospital complaints presented by the two groups are quite different, and the differences are in accordance with expectations."
Cerf (8)	2107 aviation trainees	Ratings of satisfactory-unsatisfactory adaptability, made by psychiatric interviewers	The critical ratio of the mean score difference for the satisfactory and the unsatisfactory group was 4.80. Biserial correlation of −.35, significant at .01 level.
Cerf (8)	194 WASPS	Ratings of satisfac-factory-unsatisfactory adaptability, made by psychiatric interviewers	Biserial correlation of −.36 (significant) between inventory scores and criterion ratings.
Heathers (21)	303 AAF enlisted personnel	Psychiatric patients vs. non-psychiatric patients	A critical ratio of the mean score difference of 8.7 was found. The biserial correlation between inventory stores and the criterion was .56.
Heathers (21)	202 AAF enlisted returnees	"Normals" vs. anxiety-reaction cases	A critical ratio of the mean score difference of 7.3 was found.
Mote, Berry & Graham (38); also Berry, Leavitt, and Mote (6)	491 Naval training recruits	"Normals" vs. neuropsychiatric ward cases	The inventory identified 52% of the neuropsychiatric discharges, and included 18% false positives.
Shaffer (47)	85 AAF officers	"Normals" vs. anxiety-reaction cases	Biserial correlation coefficients of .43 and .45 were found between inventory scores and the criterion.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Personal Inventory (*continued*)	
Shaffer (47)	302 AAF personnel	"Normals" vs. anxiety-reaction cases	Critical ratios (*t*'s) of the mean score group-differences were 6.75 between "normals" and mild anxiety-reaction cases, and 5.5 between "normals" and severe anxiety-reaction cases.
Shaffer (47)	199 AAF pilots	"Normals" vs. anxiety-reaction cases	A critical ratio of the mean score group-difference of 7.16 was found.
Shaffer (47)	154 AAF navigators	"Normals" vs. anxiety-reaction cases	A critical ratio of the mean score group-difference of 5.02 was found.
Shaffer (47)	172 AAF bombardiers	"Normals" vs. anxiety-reaction cases	A critical ratio of the mean score group-difference of 5.67 was found.
Shaffer (47)	994 AAF personnel	"Normals" vs. anxiety-reaction cases	A biserial correlation coefficient of .52 between inventory scores and criterion was obtained.
Shaffer (47)	802 AAF personnel	"Normals" vs. anxiety-reaction cases	A biserial correlation coefficient of .45 between inventory scores and criterion was obtained.
Shaffer (47)	859 AAF personnel	"Normals" vs. anxiety-reaction cases	"If a cut-off score of 10 or less were used, 49% of the anxiety-reaction cases but only 12% of the normals would be included. The separation is not quite so effective at the upper end of the distribution."
Shaffer (47)	2720 AAF returnee officers	"Normals" vs. anxiety-reaction cases	"Item analysis . . . showed that a large proportion (31 out of 45) of the new items were valid."
Shaffer (47)	1515 AAF personnel	"Normals" vs. anxiety-reaction cases	18 out of 20 items on the schedule were shown to be valid by item analysis.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Personal Inventory (*continued*)	
Shipley & Graham (48)	623 Naval personnel	"Normals" vs. psychiatric discharges	"Differentiation continued at a high level . . . for example, one cutting score identified 52% of the discharges while including 4% of the normal men."
Shipley & Graham (48)	571 Naval recruits & 491 limited service men	Psychiatric disposition subsequently made of the men	"Good differentiation was again found. One cutting score identified 52.5% of the psychiatric discharges, while including less than 7% of the normals."
Shipley, Gray & Newbert (49)	1385 enlisted Naval personnel	"Normals" vs. psychiatric discharges	All sixty items of the original scoring stencil continued to differentiate successfully between discharges and "normals." Critical ratios of the mean score group-differences ranged from 2.4 to 15.9 for the 60 items.
Shipley, Gray & Newbert (50)	538 Naval trainees and 263 psychiatric ward discharges	"Normals" vs. psychiatric discharges	"A critical score of 8 . . . identified 68.8% of the psychiatric discharges and included but 4.45% of the 'normals.' The validity of each item also proves satisfactory, with critical ratios ranging from 3.8 to 16.7."
Stoiurow and Schrader (55)	136 AAF enlisted men	Combat returnees vs. non-combat enlisted men	"The ex-combat men obtained higher scores on the inventory, which was indicative of greater maladjustment within this group. The difference between the mean scores of the two groups was significant . . . " (t = 2.43).
Stone and Malament* (56)	1266 Maritime Service trainees	Non-diagnosed men vs. trainees disenrolled for neuropsychiatric reasons	At a cutting score of 20, 65.4% (N = 30) of the neuropsychiatric disenrollees were detected, at a cost of 7.3% (N = 88) false positives.

* In this study the New London NDRC Inventory, a forerunner of the Personal Inventory, was employed.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Personal Inventory (*continued*)	
Wexler (**65**)	2152 Naval enlistees entering recruit training.	"Normals" vs. referrals returned to duty and neuropsychiatric discharges	With a cutting score set to refer 20% of the total group for psychiatric check, and excluding the 57 referrals who were returned to duty, there were 72% true positives, 16% false positives, and 28% false negatives; the respective *N*'s are 71, 319, and 28.
Wexler (**65**)	340 Naval enlistees (242 normal, 98 maladjusted)	Ratings by psychiatrists and psychologists	With a cutting score set to refer 30% of the total group for psychiatric check, there were 69% true positives, 14% false positives, and 31% false negatives; the respective *N*'s are 68, 34, and 30.
		Cornell Selectee Index	
Dynes (**11**)	2000 Naval personnel returned from duty	Hospitalization for nervous breakdown on basis of psychiatric examination	All the men who had to be hospitalized for nervous breakdown obtained a CSI score of 25 or more.
Harris (**20**)	Naval personnel applicants at a Navy Yard	Neuropsychiatric screening	"Those showing a score of 25 or more invariably fell into the category of severe psychoneurotics and could, therefore, be earmarked for careful questioning. Those showing scores of less than 15 could almost as readily be accepted for employment."
Heathers (**21**)	300 AAF enlisted men	Anxiety-reaction and psychoneurotic patients vs. non-psychiatric patients	Critical ratios of the mean score differences from 3.2 to 7.1 show that the CSI "gives a reliable differentiation of anxiety-reaction and psychoneurotic patients . . . from orthopedic and medical surgical patients."
Heathers (**21**)	300 AAF personnel	Anxiety-reaction and psychoneurotic patients vs. non-psychiatric patients	A critical ratio of the mean score difference of 8.2 between non-psychiatric and anxiety-reaction patients; and of 10.9 between non-psychiatric and psychoneurotic groups.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Cornell Selectee Index (*continued*)	
Manson and Grayson (33)	200 Army military prisoners	Sick-book riders vs. non-sick-book riders	"The Cornell Selectee Index is an excellent instrument for identifying 'sick book riders' and eliminating non-'sick book riders,' when the critical score of 25 or more is used. The CSI does not differentiate to any marked degree between 'sick book' and non-'sick book riders' for the 'mild psychoneurosis' or the 'non-psychoneurosis' categories."
Weider and Wechsler (61)	1000 selectees	Neuropsychiatric screening	89% ($N=89$) of the men rejected by the neuropsychiatric screening were also screened by the Index, with 12% ($N=110$) false positives.
Weider and Wechsler (61)	204 neuropsychiatric discharges and 406 men accepted for military training	Neuropsychiatric disability discharge vs. military training acceptance	82% ($N=173$) of the discharges were also detected by the Index, with 6% ($N=27$) false positives.
Weider and Wechsler (61)	600 psychiatric accepts and 400 psychiatric selectee rejects	Neuropsychiatric screening	71% ($N=284$) of the psychiatric rejects were also screened by the Index, with 15% ($N=89$) false positives, at a cutting score of 15 plus 1 stop question.
Weider and Wechsler (61)	539 "normal" selectees; 142 psychoneurotic discharges; 260 moderately psycho neurotic men; 39 mildly psychoneurotic men	Neuropsychiatric diagnosis	.9% ($N=5$) of the "normals," 13% ($N=5$) of the "mild" psychoneurotics, 78% ($N=202$) of the "moderately severe" psychoneurotics, and 92% ($N=131$) of the "severe" psychoneurotics were screened at a cutting score of 23.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Cornell Selectee Index (*continued*)	
Weinstock and Watson (64)	212 Naval training recruits	Cornell Selectee Index diagnoses vs. psychiatric discharge	Of 91 recruits with high CSI scores who were retained "on trial," only 38 (or 23%) had to be discharged during training. 91 of the 164 recruits with high CSI scores had been passed as acceptable by clinical judgment prior to the CSI; only 8 of these 91 had to be discharged.
Wexler (65)	2152 Naval enlistees entering recruit training	"Normals" vs. referrals returned to duty and neuropsychiatric discharges	With a cutting score set to refer 20% of the total group for psychiatric check, and excluding the 57 referrals who were returned to duty, there were 71% true positives, 17% false positives, and 29% false negatives; the respective *N*'s are 70, 339, and 29.
Wexler (65)	340 Naval enlistees (242 normal, 98 maladjusted)	Ratings by psychiatrists and psychologists	With a cutting score set to refer 30% of the total group for psychiatric check, there were 67% true positives, 15% false positives, and 33% false negatives; the respective *N*'s are 66, 36, and 32.
Wolff & others (75)	307 selectees	Neuropsychiatric screening	86% of the men rejected by the neuropsychiatric screening were also screened by the Index.
Wolff & others (75)	1863 selectees	Neuropsychiatric screening	80% of the men rejected by the neuropsychiatric screening were also screened by the Index.
Wolff & others (75)	1390 selectees	Neuropsychiatric screening	87% of the men rejected by the neuropsychiatric screening were also screened by the Index.
Wolff & others (75)	282 Army and Navy discharges	Discharge for neuropsychiatric disorders	88 to 90% of the men discharged were screened by the Index.
Wolff & others (75)	63 selectees who had already been screened	Second neuropsychiatric examination	87% of the men rejected by the second neuropsychiatric screening were also rejected by the Index.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Cornell Selectee Index (*continued*)	
Wolff & others (75)	50 selectees who had already been interviewed	Second neuropsychiatric examination	In four cases the second interviewer reversed the decision of the first one partly as a result of the added information afforded by the Index.
Wolff & others (75)	380 Southern white and Negro selectees	Neuropsychiatric screening	The Index screened 30 out of 31 men rejected by the neuropsychiatric interview method.
		Maritime Service Inventory	
Coville (10)	678 Maritime Service trainees	"Normals" vs. neuropsychiatric subjects	With a cutting score of 20, 66% of the neuropsychiatric subjects were identified, at a cost of 2 false positives. When "stop" items were included, 93% true positives were detected to 2% false positives. Critical ratio of the mean score difference, 21.5
Stone & Malament (56)	1208 Maritime Service trainees	Non-diagnosed men vs. trainees disenrolled for neuropsychiatric reasons	At a cutting score of 21, 66% ($N=68$) of the neuropsychiatric disenrolless were detected, at a cost of 13% ($N=144$) false positives.
Stone & Malament (56)	1018 Maritime Service trainees	Non-diagnosed men vs. trainees disenrolled for neuropsychiatric reasons	At a cutting score of 15, 60% ($N=34$) of the neuropsychiatric disenrollees were detected, at a cost of 12% ($N=111$) false positives.
Varney & Stone (59)	813 Maritime Service trainees disenrolled for neuropsychiatric reasons	Disenrollment for neuropsychiatric reasons	Of 813 disenrollees, 566 (70%) were detected by procedures in which the MSI played a prominent role; 247 (30%) were disenrolled after having passed through the screening procedures unrejected.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Minnesota Multiphasic Inventory	
Benton (4)	85 Naval neuropsychiatric patients	Neuropsychiatric ward status	Five out of 10 schizophrenics gave positive results on the Schizophrenia Scale; 5 out of 9 hysterics were positive on the Hysteria Scale; 13 of 16 delinquents were positive on the Psychopathic Deviate Scale.
Benton & Probst (5)	76 neuropsychiatric Naval patients	Ratings by psychiatrists	"In the case of the *Psychopathic Deviate*, *Paranoia*, and *Schizophrenia* trends the differences with respect to mean test score between the normal and the abnormal groups can be considered to be significant (CR's 2.6 to 3.2). . . . On the other hand, there is no substantial amount of agreement with respect to the strength of the *Hypochondriasis*, *Depression*, *Hysteria*, *Femininity*, and *Psychasthenia* trends."
Gough (17)	166 enlisted men	"Normals" vs. psychiatric hospital admissions	Significant critical ratios showed that the MMPI successfully differentiated among several diagnostic subgroupings.
Leverenz (31)	105 psychiatric ward Army cases	Neuropsychiatric diagnosis	"In the majority of the cases the scores on the Inventory did confirm the clinical impression." (No statistical data given.)
Modlin (36)	316 enlisted Army personnel	Neuropsychiatric ward cases vs. "normals"	"Depression was most successfully verified by the Multiphasic Inventory, inasmuch as 88% of 31 clearly classified depressives scored highest on the *D* Scale . . . A close correlation with clinical expectations is seen in most of the categories."

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Minnesota Multipharic Inventory (*continued*)	
Morris (37)	320 Naval personnel under psychiatric observation	Neuropsychiatric diagnosis	"The nosological groups under consideration could not be differentiated from one another on the basis of Inventory scores . . . The Inventory does differentiate borderline normals from serious pathological states but does not aid in the differential diagnosis among the pathological groups."
Schmidt (46)	211 AAF personnel	"Normals" vs. men diagnosed as neuropsychiatric cases	Critical ratios of the mean score group-differences showed significant inventory differences bebetween "normals" and men diagnosed as psychopaths, neurotics, and psychotics.
		Experience Comparison Index	
Owens (39); and Owens and Zirkle (40)	400 enlisted Naval personnel	"Normals" vs. men diagnosed psychiatrically as maladjusted	"False positive scores not only ranged as high as 30, but they occurred more frequently than the true positive scores." 187 positive to 217 false positives out of 400 tested men.
Wexler (65)	582 Naval enlistees (209 well-adjusted, 241 doubtful, 132 poorly adjusted)	Ratings by psychiatrists and psychologists	With a cutting score set to refer 40% of the total group for psychiatric check, and *excluding the 241 doubtful cases*, there were 71% true positives, 13% false positives, and 29% false negatives; the respective *N*'s are 94, 27, and 38.
		Personal Check List	
Owens (39)	600 enlisted Naval personnel	"Normals" vs. men psychiatrically diagnosed as maladjusted	The Personal Check List referred 417 positives to 183 false positives.
Wexler (65)	561 Naval enlistees (418 normal, 143 maladjusted)	Ratings by psychiatrists and psychologists	With a cutting score set to refer 30% of the total group for psychiatric check, there were 75% true positives, 15% false positives, and 25% false negatives; the respective *N*'s are 107, 63, and 36.

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Billet Qualifications Blank	
Wexler (65)	2152 Naval enlistees entering recruit training	"Normals" vs. referrals returned to duty and neuropsychiatric discharges	With a cutting score set to refer 20% of the total group for psychiatric check, and excluding the 57 referrals who were returned to duty, there were 69% true positives, 17% false positives, and 31% false negatives; the respective *N*'s are 68, 339, and 31.
Wexler (65)	268 Naval enlistees (104 well adjusted, 112 doubtful, 52 poorly adjusted)	Ratings by psychiatrists and psychologists	With a cutting score set to refer 40% of the total group for psychiatric check, and *excluding the 112 doubtful cases*, there were 78 true positives, 13% false positives, and 22% false negatives; the respective *N*'s are 41, 14, and 11.
		Convalescent Personal Inventory	
Heathers (21)	235 AAF enlisted personnel	Anxiety-reaction patients vs. non-anxiety reaction patients	A critical ratio of 9.2 "indicates a high degree of differentiation."
Heathers (21)	441 AAF enlisted personnel	Psychiatric vs. non-psychiatric patients	A biserial correlation of .38 was found between the inventory scores and the criterion. "The psychiatric group was found to have a reliably greater number of significant responses."
		Neuropsychiatric Adjunct Inventory	
H. C. Leavitt (26)	768 military inductees	Ratings by psychiatrists	About 85% of the men presenting the more common psychopathological syndromes were successfully screened by the inventory.
		Questionnaire Regarding Present Reactions	
Heathers (21)	200 AAF enlisted personnel	Psychiatric vs. non-psychiatric patients	A critical ratio of the mean score difference of 6.2 was found; and a biserial correlation of .67. "A critical score of 40 significant responses screened 69% of patients diagnosed as psychiatric and 21% of patients diagnosed as non-psychiatric."

TABLE I (*continued*)

Source	*Group Tested*	*Criterion*	*Result*
		Inventory of Psychological Problems	
Heathers (21)	426 AAF enlisted personnel	Psychiatric vs. non-psychiatric patients	Critical ratios of the mean score differences between psychiatric and non-psychiatric groups ranged from 1.4 to 7.6. "The total scores on frequency and severity items of the inventory discriminated psychiatric and non-psychiatric groups of patients.
		Officer Personal Inventory	
Smith & Voss (53)	1383 officers	Qualified officers vs. emotionally disqualified officers	"A cut-off score of 30 on this key marked off 53.01% of the emotionally disqualified officers and 2.62% of qualified officers." Significant critical ratios were obtained for all the items of the test.
		Psychoneurotic Inventory	
Page (41)	244 Army trainees	Number of times the men went on sick call	Coefficient of correlation between the number of times the men went on sick call and their Inventory scores was .25.
Page (41)	600 enlisted Army personnel	Neurotics vs. undiagnosed men	The critical ratio of the mean score difference between groups was 15.7; the biserial correlation coefficient was .78.
		Bernreuter Personality Inventory	
Page (41)	200 enlisted Army personnel	Neurotics vs. undiagnosed men	The critical ratio of the mean score difference between groups was 12.27; the biserial correlation coefficient was .75.
		Psychosomatic Inventory	
Page (41)	200 enlisted Army personnel	Neurotics vs. undiagnosed men	The critical ratio of the mean score difference between groups was 8.8; the biserial correlation coefficient was .80.

The remarks of Smith and Voss, reporting on their validation of the Officer Personal Inventory, are of interest in this connection:

> The second restrictive condition was the impossibility of establishing complete independence between the Officer Inventory scores and the diagnoses of emotional fitness, at the time the diagnoses were made. In the established classification procedure, which it was not possible to alter, the Boards of Medical Examiners had available at the time of their examinations, total Officer Personal Inventory scores, based on a tentative 34-item scoring key, for over half of the officers concerned in this study. Hence, there may have been some suggestive effect exercised by the knowledge of the Inventory scores upon the diagnoses of emotional fitness . . . (53, pp. 1–2).

In other words: even when military experimenters were well aware of the necessity of avoiding criterion contamination, military conditions sometimes made it impossible for them to eliminate all such contamination. Consequently, several of the obtained validity coefficients are doubtless higher than they should be. The exact extent of validity-inflation, however, is extremely difficult to evaluate.

2. *Criterion overlap.* Aside from the possibility of criterion contamination, overlap between criterion and inventory seems to have occurred in some of the validity investigations. Thus, even when the psychiatrists' judgments of the respondents were made wholly independently of these respondents' inventory answers, the questions asked by the psychiatrists in several instances duplicated those included in the inventories (or vice versa). This means that the prognosis of the psychiatrists and the inventories might be in close agreement, without either one's being necessarily accurate in terms of the actual outcome of the prognosticated cases. The obtained validity coefficients might consequently be spuriously high.

Overlap between the criterion and the inventory under investigation may be minimized by employing an *actual outcome* rather than a mere *prognosis.* Thus, the criterion may be actual success or failure to adjust to military life, or it may be neuropsychiatric discharge because of breakdown in military service. But in many of the reported studies the criteria could not be of this definitive nature. Often the criteria were merely prognoses based on rather brief psychiatric interviews. In these latter instances, criterion overlap was common, and is judged to have caused validity coefficients which must be taken with at least "a grain of salt."

3. *Sample heterogeneity.* The unscreened recruits tested by the armed forces were frequently extremely heterogeneous, including at the lower end unemployables, tramps, loafers, "bums," alcoholics, frank neurotics and so on (9). It should be relatively easy (by almost any technique) to sort out such individuals. The samples in civilian studies of personality inventories, on the other hand, have usually included relatively few of such easily diagnosable cases. This probably accounts, at least in

part, for the higher degree of success in the military reports on personality inventories.

4. *Extreme or atypical validation groups.* In several of the military studies of personality inventories, the groups that were employed to test the validity of the inventory were of a somewhat special or extreme sort. Of course the "method of extreme groups" is of recognized value in the *development* of a test; but for a check on practical, operating validity, application of the test instrument to an unselected sample appears preferable. If, for example, a sample contains few or no "doubtful" cases, it is free of the very cases which are hardest to diagnose. Similarly, if a sample contains an unusually large proportion of maladjusted cases, the proportion of "false positives" to true positives at any given cut-off score will necessarily be lower than in a sample containing only a normal complement of maladjusted. (If the sample contained *only* maladjusted cases, there could not be *any* false positives.) Such facts have not always received full recognition in validity-studies of screening instruments. Some studies in which samples of abnormal composition were used, with inferences as to validity, are cited below.

In Page's (41) and Heathers' (21) studies, groups were used in which the number of abnormal cases equalled the number of normal—clearly an atypical ratio.

In some of Wexler's (65) tables (though not in his graphs), the percentage of false positives has been figured on a group from whom "doubtful" cases have been excluded. (In one sample of 582 men, the "doubtful" cases constituted no less than 41% of the total.)

Smith and Voss (53) found that the Officer Personal Inventory distinguished significantly between 1300 qualified officers and 83 emotionally disqualified ones. But the qualified men had already been passed by a medical examiners' board, and were therefore (so to speak) "super-normal" rather than "normal." Thus the sample presumably included relatively few of the borderline group—differentiation of whom is ordinarily the most difficult task of an inventory.

The "normal" group used in Schmidt's (46) study obtained Minnesota Multiphasic Inventory mean scores of below 50 on eight of the nine MMPI scales—thus indicating that it, too, was probably super-normal.

Shipley, Gray, and Newbert (49), in a study involving the Personal Inventory, reported that "the normal group comprised 1004 newly enlisted men who had been favorably passed upon in the psychiatric interview; the deviating group, 385 early psychiatric discharges tested while under observation, and prior to discharge, on the psychiatric ward" (49, p. 2). Here again a screened or "super-normal" group is being compared with a definitely abnormal one. Moreover, the ratio of 385 "early psychiatric discharges" to 1004 is clearly atypical: a more normal ratio would appear to be only 50 to 1000, or perhaps as high as 100 to 1000 (65, p. 126).

Regarding military studies which make use of hospitalized patients, Wexler makes the following pertinent observation:

Nothing is so apt to produce delusions of grandeur in the constructor of

personality inventories than the use of what might be termed the "ex post facto" psychiatric criterion. This involves a technique by which a hospitalized population is contrasted with successful military personnel in a training center. The difficulty with this procedure is that patients in a neuropsychiatric ward have generally become so sensitized to clinical symptomatology that they no longer react in the way in which they would have reacted to the same instrument prior to hospitalization. Furthermore, the hospital wards generally contain a selected sample from the extreme of the deviated cases. This means an extremely stringent criterion group. . . . It is no remarkable thing, then, that test score differentiation from a young, enthusiastic, and healthy recruit group can be obtained with considerable ease (65, pp. 140–141).

5. *Differential motivation.* It seems likely that in some of the armed forces' validation studies the "abnormal" criterion groups were differently motivated than the "normal" groups; and that consequently some of the obtained significant differences were exaggerated. Relevant here are the following considerations.

a. A good many of the neuropsychiatric groups that were used as criterion groups consisted exclusively of neuropsychiatric ward cases. It is possible that these ward cases, who were only a step or two from transfer to inactive service or from complete release, had a definite incentive to answer the inventory questions in a self-incriminating manner. This would clearly differentiate them from non-hospitalized "normals," and spuriously boost the obtained validity coefficients.

b. In the study by Manson and Grayson (33) employing the Cornell Selectee Index, it was found that this Index differentiated significantly between sick-book riders and non-sick-book riders. It is stated by the authors that the sick-book riders were not discouraged from obtaining temporary escape from the rigorous training program of the Center where they were being held as prisoners. Thus they were allowed to achieve a definite "neurotic gain" from their sick-book riding. But since the CSI consists largely of psychosomatic questions; and since the sick-book riders (who were tested *after* they became sick-book riders) were already on record as having various psychosomatic complaints—it would be odd if they did other than check considerably more CSI responses than did non-sick-book riders.

c. Altus and Bell, after reporting significant inventory score-differences between successful and failing illiterate Army Special Training Center inductees, are careful to note that "one *caveat*, however, must be entered. In the ordinary school situation, passing a test or successfully completing a course has value as a goal toward which to strive. For some, if not many, of the trainees, passing the tests for graduation may not represent a desirable goal; for it means retention in the Army, an outcome which occasionally finds some quite intelligent soldiers unenthusiastic. Becoming a soldier meant to the average trainee giving up a job which paid him two to three times more than he ever earned in his

life before" (2, p. 103). In other words: some of the trainees (whether consciously or unconsciously) had a definite motive (*a*) for failing the inventory—that is, appearing to be neurotic; and (*b*) for failing the course. The observed significant inventory-differences, therefore, between those who passed and those who failed, are probably partly spurious.

d. In separate studies by Gough (17) and Schmidt (46) it was found that such diagnosed groups of military personnel as psychopaths, neurotics, and psychotics obtained Minnesota Multiphasic Personality Inventory scores that significantly differentiated them from the "normal" controls on nearly *all* the MMPI scales. This would seem to indicate that either the MMPI scales do not actually differ from each other as they are supposed to; or that the respondents in the psychiatric groups were motivated deliberately to answer *all* kinds of MMPI questions in a manner unfavorable to themselves. Curiously enough, in the Gough study the *Lie* scores of the psychiatric respondents were always in the normal range, but the *?* scores of the psychiatric subjects were almost always considerably below those of the "normals." It may therefore be wondered if the psychiatric respondents were able to avoid the *Lie* questions (which are fairly transparent) but were so eager to commit themselves to unfavorable answers to the other items that they gave suspiciously few *?* responses.

e. In one of the investigations reported by Shipley and Graham (48), it was found that, for some unexplained reason, 2301 Amphibious Forces men, who had already been psychiatrically screened and designated as "normal" cases, obtained a mean Personality Inventory score of 15.2—which was definitely worse than the mean score of 14.6 obtained by 74 psychiatrically disapproved Submarine School personnel. Such a reversal, when obtained with an inventory that was generally successful in differentiating between psychiatric cases and "normals," suggests first that the psychiatric standards applied to the submarine men were higher than those in the Amphibious Forces. Another possibility is that the Amphibious Forces men were indifferent to whether or not they made favorable scores on the Inventory, while the submarine men tried to make as good a showing as possible. In this connection, a comment by Wexler is pertinent:

> Men applying for submarine duty are highly motivated. As a volunteer group, they are extremely eager to make a good impression and qualify for submarine training. Raw recruits, too, generally have reasonably high motivation since all the indications are that men passing through the initial training stages are generally eager to do well in the service, though the motivation is hardly as high as in the case of the submarine group. There is reason to believe that the amphibious forces were somewhat less motivated. Certainly the motivation of the combat-experienced groups, at least as far as the tests were concerned, was less than for any of the other groups. Simple observation has confirmed the fact

that these men were indifferent to the tests and cynical concerning anything except the chance for immediate leave (65, p. 168).

In sum, it seems clear that differences in group-motivation can cause significant differences in personality inventory scores. This factor requires attention in evaluating observed group-differences.

6. *Honesty of response.* The preceding section has cited instances where responses in certain groups were probably distorted either in the direction of self-incrimination (e.g., neuropsychiatric ward cases) or in the direction of self-inflation (submarine men). There is, however, the possibility that, on the average, the members of the armed services answered the personality inventories with less distortion than civilians ordinarily do; and that, in consequence, the inventories tended to be more effective in military than in civilian practice. Relevant in this connection are the following points.

a. Military personnel may be specifically encouraged and warned to give honest inventory responses by test administration techniques that are impossible with civilians. Thus, in Coville's study (**10**) the subjects were warned that dishonest answers would inevitably be discovered later and would then lead to a dishonorable discharge. They were also cautioned that, for their own medical well-being, they should be particularly honest about admitting physical and nervous symptoms.

b. Harris, using the Cornell Selectee Index with Naval personnel, makes the following report:

> Rarely do patients fail to answer truthfully the question, "Were you ever a patient at a mental hospital?" On the contrary, they have usually been so anxious not to be caught in written untruths, that they are likely to answer "Yes" to that question, whereas, actually, the confinement was to a general hospital for injury or nonpsychiatric observation. Likewise, the question, "Have you ever had a fit or a convulsion?" is also answered with meticulous truthfulness, thereby leading to a diagnosis of epilepsy or hysterical loss of consciousness, equally undesirable in Yard employees or Navy personnel. Often, in his anxiety to answer the question correctly, the applicant is led to include such occurrences as "fits of temper" or simple syncope, also important in his evaluation (20, p. 596).

c. Wolff and his associates, after considerable experience with the Cornell Selectee Index, note that "although responses may be falsified, in practice we have noted this infrequently" (**75**, p. 9). Civilian studies, on the other hand, show quite frequent and consistent falsification or exaggeration of personality inventory scores (**12**, pp. 414–420).

d. A discordant note is struck by Cerf[8] who, employing the Information Blank, S-C, CE 410 A with Army Air Force personnel, discovered that "the truthfulness scores of the group were generally low. Of the ten truthfulness items, the average aviation student made truthful re-

[8] For a summary of Cerf's study, see the 22nd entry in Table II.

sponses to only 5.4." This indicates "a substantial amount of falsification of response" (8, p. 580). It is perhaps significant that in this particular study the inventory employed did not successfully distinguish between passing and failing aviation students.

7. *Role of intelligence.* In some of the military studies of personality inventories, the poor showing of some respondents may have been due to their failure fully to comprehend either the instructions of the inventory, the content, or both. If (as seems rather likely) such respondents tended also to be relatively unadaptable to military life, the effect would be to boost the military validity of the personality inventories. In this connection, Shipley, Gray, and Newbert noted a "tendency . . . for men with low General Classification Test scores to show up as maladjusted on the Personal Inventory" (51, p. 7). This observation is based on a correlation of −.28 between AGCT scores and Personal Inventory scores of the testees. Such a correlation is too low to give much support to the idea of a common intelligence-factor in inventory-scores and adaptability to military life; but it is quite possible that the relation is curvilinear, and stronger in the lower reaches of intelligence than in the levels above.

Altus and Bell, administering their inventory orally to Army illiterates, found that the inventory scores indicated " . . . considerably more hysteria, hypochondria, paranoia, and depression among these men of low socio-economic status and of marginal intellect than is true of groups approximating normal intelligence and socio-economic level. For these latter groups, more subtle questions would doubtless be required" (2, p. 476). If Altus and Bell are right in their remark about the need for more subtle questions, it appears that the comparatively direct questions of personality inventories probably have special validity when used with less intelligent or more naïve individuals. This may well be one of the reasons for the superior validity of the inventories in military applications; since civilian applications of the inventories have commonly been to groups that are more intelligent, better educated, and more sophisticated than the typical (enlisted) military sample.

8. *Statistical inadequacies.* In some validations of personality inventories in military practice, there were statistical inadequacies which throw doubt upon the accuracy of the obtained significant differences or validity coefficients. Some of these statistical inadequacies will now be discussed.

a. In several of the studies involving biserial coefficients of correlation, the coefficients were calculated on the basis of an almost equal number of "normals" and abnormals. Actually, since "normals" are generally far more numerous than abnormals, biserial correlation coefficients should be calculated only on groupings which approximate the "normal"-abnormal ratio. Yet in Page's (41) study several of the biserial correlations were based on 100 neurotics and 100 undiagnosed cases; and in several of Heathers' (21) validation experiments, biserial

correlations were based on 145 psychiatric and 158 non-psychiatric patients, on 100 psychiatric and 100 non-psychiatric patients, and on 215 psychiatric and 226 non-psychiatric cases. Biserial coefficients computed for such samples are spuriously high.

b. In several of the studies results are reported only in terms of the critical ratio of the mean score difference between psychiatric and non-psychiatric groups, rather than in terms of the correlation coefficient. Critical ratios, however, can be misleading when used for test validation, as shown by the instance where Williams, Leavitt, and Mendola (**73**) found a critical ratio of 3.6 for the difference between the mean scores of passing and failing Marine Corps Officer candidates who took the Personal Inventory. When the biserial correlation was calculated for the same groups, the correlation coefficient was found to be only .18. Careful analysis led the authors to conclude that "it is clear that the degree of overlapping of the *pass* and *fail* distributions is so great as virtually to preclude use of the stencil in individual selection. Only at the very extreme of the distribution would the test score be very significantly better than chance as a predictor of success or failure in Platoon Commanders School" (**73**, p. 7).

c. When correlation coefficients of validity were calculated, some of the authors of the reported studies read considerable significance into coefficients which were hardly very high. Thus, Satter, after obtaining a validity coefficient of .39 when using the Personal Inventory on parachute trainees, remarked that "in terms of what we know about the predictive efficiency of personality measures in general, this coefficient is gratifyingly high" (**44**, p. 30). While it is true that a correlation of .39 is ***comparatively*** high for personality inventory validation, it still leaves much to be desired.

d. Several of the experimenters unfortunately employed the same group for validation purposes that they had originally employed for the standardization of their instruments. This procedure will almost invariably result in spuriously high validity coefficients or critical ratios (particularly if the original sample is not very large). Smith and Voss (**53**), Page (**41**), Williams and Leavitt (**68**), and Heathers (**21**) all seem to have at times employed this particular technique. As Williams and Leavitt point out, this does not imply that the investigators were unaware of the problem; but that the exigencies of military practice sometimes precluded their using more satisfactory validation procedures. Nevertheless, the fact remains that in those instances where validation was accomplished without a fresh sample, the observed critical ratios or validity coefficients are generally too high.

e. In some of the studies, very small numbers of cases were employed in the diagnostic subgroups. Thus, in Gough's (**17**) and in Schmidt's (**46**) studies the number of individuals in the separate diagnostic categories is frequently 12 or less; and in Bobbitt and Newman's (**7**) study

of 99 U. S. Coast Guard trainees, the number of individuals in the separate categories of medical treatments is often less than 10. The present authors doubt the dependability of statistics (including *t*) based on such small samples. It is particularly in connection with small samples that the question of the differential frequency of report of positive vs. negative findings becomes important.

In sum: largely because of the exigencies of the military situations in which the personality inventory validations were made, inadequate statistical procedures were sometimes employed. In consequence, some of the obtained validity coefficients or critical ratios are questionable, and others are definitely too high.

9. *Evaluation of "false positive" results.* Because of the limited purposes to which personality questionnaires were usually applied in military practice, and because of the fairly abundant manpower available for military activity in World War II, results were sometimes accepted as satisfactory which by civilian standards might be regarded as only questionably satisfactory. Thus, large numbers of false positives were apparently accepted without very great concern, on the assumption either that the error of classification would be corrected in later checkups (usually by psychiatric interview); or that a fairly large amount of this type of error may be tolerated in military practice (so long as additional recruits can be readily obtained through the draft). A false positive diagnosis in civilian practice, on the other hand, is generally taken more seriously.

It requires note, moreover, that in the military studies false positives were usually presented in terms of *percentages* rather than *numbers of cases;* and this tends to make the false positives seem less prominent than they actually are. In actual practice, using stipulated cutting scores, it appears that roughly from 20 to 30% false positives were found among unselected recruits, while 70 to 80% of the true positives might be detected by the inventories. But since the "normal" testees usually were far greater in number than the "maladjusted" ones, 20% false positives may easily represent a greater number of men than may 80% true positives. Thus, in a study by Wexler (**65**) of Naval recruits entering training, 319 false positives were screened, to only 71 true positives; in a study by Stone and Malament (**56**), 248 false positives to 79 true positives; in a study by Satter (**44**), 211 false positives to 104 true positives; and in a study by Williams, Leavitt, and Mendola (**73**), 154 false positives to 56 true positives. This means that even where, in terms of *percentage*, the false-positive ratio is considerably smaller than the true-positive ratio, in terms of *number of men*, the group falsely classified as positive by the inventory may be considerably larger than the group correctly classified. This seems to be generally true, except in samples containing unusually large proportions of abnormal cases. Both the

percentage and number of false positives could be reduced by raising the cutting score on the inventory. But this would increase the number of false negatives (see below).

10. *Neglect of "false negative" results.* False negatives on personality inventories represent those individuals who are ***not*** detected by their inventory scores, but who later prove to be definitely maladjusted. Unlike false positives (individuals who obtain unfavorable inventory scores but who subsequently prove to be "normal"), false negatives are exceptionally disadvantageous in a military situation. For it is the false negatives who become inefficient soldiers (whether enlisted men or officers), who cause considerable trouble to others, who may jeopardize the life or success of a combat unit, who eventually have to be discharged for neuropsychiatric reasons, and who tend to become governmental charges, in one form or another, for decades after a given war has ended.

In most of the reported validity-studies of personality inventories, the numbers and percentages of false negatives are not recorded. This arises from the fact that most of these studies employed only psychiatric screening (or prognostic) criteria; and when such criteria were employed, it was customary to submit to psychiatric interview only those diagnosed as "positive" by the inventory. The "negatives" were not checked, and thus the false negatives or "misses" not detected.[9] In studies such as those reported by Varney and Stone (59), Wexler (65), or Satter (44), where the criterion was that of successful *performance* or *adaptability* to military service, the number and percentage of false negatives are more likely to be either directly reported, or readily ascertainable from the data presented.

The ratio of false negatives to true positives, where reported, seems to have been far from negligible. Theoretically, the ratio might be expected to be high (*a*) when the cutting score on the inventory is set high[10] (reducing the number of false positives at the cost of increasing the false negatives), and (*b*) when the motivation for "covering up" unfavorable responses is strong (as seems to have been the case, for example, with individuals seeking admission to the Merchant Marine, the submarine service, and officer candidate school). With a high cut-off point, Shipley and Graham—who have made the most extensive studies in this field—estimate that "by using the short form of the [Personal] Inventory, 50 percent of the potentially dischargeable men would be identified by interviewing 8 percent of the total population" (48, p. 5); the remaining 50 percent of the potentially dischargeable men constitute the false

[9] The neglect of the inventory-"negatives" was due to the extreme shortage of psychiatrists, who were kept fully occupied examining the neuropsychiatrically more likely group of inventory-"positives."

[10] A "high" score on inventories of the kind considered in this paper is one that denotes a comparatively high likelihood of neuropsychiatric unfitness for military service; the higher the score, the greater the presumed neuropsychiatric vulnerability.

negatives. While a reduction of the interview-sample to but 8 percent of the total population undoubtedly represents a precious saving of interview-time in military selection, it may be doubted whether an instrument with so high a false-negative rate would, in civilian practice, be considered of great service. Various other investigators, using lower cutting scores, have reported lower (but still very appreciable) percentages of false negatives. Thus, from data presented by Satter (**44**, p. 11), it appears that at a cutting score set to eliminate 33% of the sample, 104 of 183 recruits who failed in Parachute School (for such reasons as refusal to jump, and indifference to continuing training) would be detected, while 79 would not be; here the proportion of false negatives is 43%. From figures presented by Wexler (**65**) for a large sample of Naval recruits entering training, it is clear that, with a cutting score rejecting 20%, the proportion of false negatives is 28% for the Personal Inventory, and 29% for the Cornell Selectee Index. Varney and Stone, using an unspecified cutting score on the Maritime Service Inventory, together with a check-list of illnesses and some follow-up observation of selected cases, report that "about one-quarter of those disenrolled for neurotic and psychopathic conditions were eventually referred for attention after having passed through screening undetected" (**59**, p. 46); that is to say, in this instance the false-negative rate was about 25%. Weider and his associates (**35**, **61**, **62**, **63**), using the Cornell Selectee Index with a presumably low cutting score, report a false-negative rate of only 10–20%; this rate, while gratifyingly low, necessarily entails a compensatingly high false-positive rate. A similar remark applies to H. C. Leavitt's (**26**) findings with the Neuropsychiatric Screening Adjunct Inventory.

The personality inventories used by the armed forces also turned up false negatives in another sense; that is, by proving quite ineffective in the diagnosis of certain psychiatric syndromes while they more effectively diagnosed others. Thus, Varney and Stone state that the Maritime Service Inventory failed to screen "most psychotics [and] a large proportion of organic cases . . . " (**59**, p. 46). Similarly, H. C. Leavitt states that "a number of psychopathological syndromes are not effectively detected by the tests" (**26**, p. 356).

A practical fact in connection with false negatives is this: the veterans' hospitals are now caring for large numbers of neuropsychiatric patients, who obviously were *not* screened and culled successfully by personality inventories or psychiatric interviews. One reason for this may be that the armed services failed to apply the screening technique regularly, or to abide by the results consistently. Another reason may be that it was considered advisable to reduce the number of false positives, even at the cost of increasing the number of false negatives; if so, this of course represents an administrative adjustment to the uncertainties of the diagnostic technique; and the resulting increase in false negatives

(some of whom doubtless ended in veterans' hospitals) is chargeable to the inadequacies of the diagnostic technique.[11]

It sometimes appears that writers gave undue emphasis to the detected psychiatric cases (the true positives) rather than to the undetected ones (the false negatives). After all, the detected cases are probably those who are the least difficult to catch—as evidenced by the fact that they *were* detected. (As in crime: the criminals caught are those who are easiest to catch.)

A critic has pointed out to the writers that civilian, as well as military, validity studies tend to pay relatively little attention to the false negatives—and we agree. However, the false negatives seem to us more important in the military situation than in the civilian. In the first place, as already mentioned, the life and success of a combat unit may well hang on the teamwork and dependability of the individuals in the unit: the front line is a poor place for the malfunctioning or breakdown of a false-negative case. In the second place, the military samples to whom personality inventories are commonly applied contain a considerably larger proportion of normal people than the civilian samples to whom personality inventories are ordinarily applied in practice. In a sample with a preponderance of normals, the major error (so far as numbers of cases are concerned) would be to mis-diagnose or mis-prognose the normals as abnormal. Hence, in such a sample, it is important to reduce the proportion of false positives: and this can be accomplished by setting the cutting score high. But the effect of a high cutting score is to *increase the incidence of false negatives.* Until the inventories become more efficient, this is the necessary cost at which an acceptably low rate of false positives can be achieved. If there is any merit in either of the arguments in this paragraph, it appears that the neglect of false negatives in the military studies leads to a more optimistic conclusion than would a similar neglect in studies of typical civilian clinical groups.

11. *Specialized design, validation, and standardization.* Probably one

[11] Dr. Glenn V. Ramsey of Princeton University is inclined to take issue with this statement, on the ground that it is equivalent to demanding that an inventory predict all future neuropsychiatric breakdowns. In his own terms: "In the first place it is demanding that the inventory perform a task which is impossible even with the use of all available scientific knowledge and skills which are adaptable to a military setting. It would be more profitable and practical to attack this diagnostic problem by constructing inventories designed for specified purposes, such as the general recruit screening, adjustment to a specific program or activity, etc. Secondly, it is necessary to consider the possibility of *specificity of vulnerability*, in addition to the basic or general stability factor. The individual may or may not encounter during his military career specific situations or conditions which would precipitate a neuropsychiatric breakdown as a consequence of specific vulnerability. Demanding that an inventory predict breakdowns from both general and specific factors along with those resulting from psychosis, brain pathologies, epilepsy, etc., is asking for a degree of efficiency that is beyond reasonable expectations, in view of our present knowledge concerning these matters."

of the main reasons for the relative success of personality inventories in military practice is the fact that in many instances the instrument was specifically designed for the group to whom it was ultimately applied, and specifically validated and standardized in this same group. Further, the tests which were most frequently employed in military practice—such as the Personal Inventory and the Cornell Selectee Index—were not only originally designed for military men and their problems; but also, when these instruments were applied to specialized military groups (such as Air Force candidates or submarine trainees) they were frequently modified for the particular groups to whom they were applied.

For example, Shipley and his associates (**48**, **49**, **52**), when validating the Personal Inventory, consistently determined item validities which applied specifically to each new kind of group. Almost all the other military experimenters who employed the Personal Inventory also tended to follow this custom-tailored validation procedure. Wolff and his associates, reporting on the standardization of the Cornell Selectee Index, noted that "it is worthy of emphasis that each item in the Index was incorporated only after having been exposed to an exhaustive item analysis and statistical validation" (**75**, p. 3). Here again this type of specialized validation procedure paid good dividends. Altus and Bell (**2**), in working with illiterate Army Special Training Center candidates, carried out item validations on the Bell, Minnesota Multiphasic, and Army Adjustment schedules before they selected the particular questions which applied to the group with whom they were working; and then they devised an oral administration of their composite inventory, since the ordinary printed form would obviously be useless with illiterates.

The military experimenters, moreover, seem always to have employed some specific criterion group in their standardization procedures. They did *not* follow the too common civilian practice of "validating" their questions exclusively in terms of the criterion of internal consistency. If they wished to measure neurotic or psychotic tendencies with their instruments, they invariably employed neurotic or psychotic criterion groups in their standardization procedures. This kind of specialized external validation almost surely accounted for much of the success of the inventories employed by the armed forces.

In defense of civilian practice, it must be mentioned that the specialized construction, validation, and standardization of personality inventories requires large funds and large, appropriate samples—conditions which the civilian experimenter can rarely hope to achieve.

12. *Realistic application.* While civilian psychologists and educators sometimes apply personality inventories to diagnostic problems for which they were not originally intended, the military users of these instruments were usually much more realistic, and demanded of the tests only the limited applications for which the tests are suited. As Zubin has aptly stated:

Perhaps the most important factor was the lower level of aspiration which these inventories adopted. They were, from their very beginning in World War

I, not regarded as personality tests; they were merely sieves separating the recruits into two groups—those who had to be screened further by a clinician in a short personal interview, and those who needed no further screening" (76, p. 58. See also reference 24).

This, apparently, is what personality inventories *can* do; and this, and little more, is what their military users commonly asked of them.

Military Validation Making Use of Performance-Measures

The second method of estimating the value of personality inventories in military practice made use of performance-measures. Thus, inventories were commonly administered to the members of a training course, and later the relation was determined between inventory scores on the one hand, and success or failure in the course, on the other. The relevant data from military studies using some kind of performance-measure are listed in Table II.

Of the studies summarized in Table II, it may be observed that only seldom did the personality inventories prove distinctly effective in predicting or discriminating successful from unsuccessful performance. This negative finding suggests that the role of intelligence in favoring both successful performance *and* "adjusted" responses to the inventory-items is not very strong. The negative finding is, of course, quite the opposite of the generally favorable results when the inventories were validated against psychiatric criteria. Some reasons for this contrast are suggested below.

1. *Prior elimination of abnormals.* In many of the military studies, both the "pass" and the "fail" groups had undergone some prior selection with respect to minimum emotional fitness. Hence the lower end

TABLE II

Military Validity Studies of Personality Inventories, Making Use of Performance-Measures

Source	*Group Tested*	*Performance-Measure*	*Result*
		Personal Inventory	
Cerf (8)	1419 AAF pilots in primary training	Pass vs. fail	A biserial validity coefficient of .06 was found. "The obtained coefficient is barely significant at the 5% level, but it is in the unexpected direction. . . . The Shipley Personal Inventory, format B, is not useful as an instrument for the prediction of graduation-elimination from primary pilot training."

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		Personal Inventory (*continued*)	
Graham, Mote, & Berry (18)	2608 submarine school trainees	Pass vs. fail on tank escape performance	A tetrachoric validity coefficient of .14 was found; and a non-significant Chi-square of .58 for 1 d.f. was found between pass-fail criterion and *PI* scores.
Leavitt, Williams, & Lipkin (29)	670 Marine officer candidates	Pass vs. fail	The obtained biserial validity coefficient of .28 "is not sufficiently high to warrant use of the MFRL Stencil in individual prediction, but it is high enough for limited selection of large groups."
Lepley (30)	121 AAF radar trainees	Pass vs. fail	Non-significant validity coefficients ranged from −.15 to .06 on various subtests.
Lepley (30)	297 AAF lead bombardiers	Strike-photo analysis: radial errors in 100's of feet	Non-significant biserial validity coefficients ranged from −.07 to .10 on various subtests.
Lepley (30)	303 AAF lead bombardiers	Strike-photo analysis: per cent hits within 1000 feet	Non-significant biserial validity coefficients ranged from −.12 to .07 on the subtests.
Lepley (30)	267 AAF lead navigators	Strike-photo analysis: radial errors in 100's of feet	Non-significant biserial validity coefficients ranged from −.08 to .07 on the subtests.
Lepley (30)	265 AAF lead navigators	Strike-photo analysis: per cent hits within 1000 feet	Non-significant biserial validity coefficients ranged from −.09 to .11 on the subtests.
Satter (45)	1400 enlisted Naval personnel	Ratings of the men by their submarine officers	Correlations between Personal Inventory scores and the criterion ratings were low, and in no case significantly different from zero.
Satter (44)	1079 enlisted parachute trainees	Pass vs. fail	A validity correlation of .39 was obtained between the pass-fail criterion and the inventory scores. Critical ratios (*t*'s) of the mean score group-differences were always 4.1 or greater.

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		Personal Inventory (*continued*)	
Shipley, Gray & Newbert (51)	1466 Naval trainees	Pass vs. fail	"The Personal Inventory identified a significant proportion of the 52 men who were later discharged. 21% of these had received scores of 18 or above on the *PI*, as compared with but 5% of the active men. The mean score for the discharges was significantly higher—4.3 points (CR = 3.9)—than that for the active group."
Williams & Leavitt (68)	1039 officer candidates	Pass vs. fail	"The Personal Inventory Form 4 successfully identified a significant proportion of the men who later failed in OCS . . . . The extent of this relation is indicated by a correlation of .48 between *PI* scores and OCS success and failure (computed biserially)."
Williams, Leavitt, and Blair, (71); also Leavitt & Adler (27)	185 Marine officer candidates	Ratings by superior officers on combat proficiency	"There is very little, if any, relationship between the inventory scores and the combat rating . . . . At no point on the *PI* scale could a cutting score have been selected which would profitably eliminate men who would later prove to be poor in combat . . . . The tetrachoric correlation turns out to be .15, an insignificant figure."
Williams, Leavitt, & Mendola (73)	757 Marine Corps officer candidates	Pass vs. fail	The critical ratio of the mean score group-difference was 3.6; and the biserial validity coefficient was .18.
		Personal Inventory and Cornell Selectee Index	
Stolurow, Irion, & Pascal (54)	300 instructors at AAF Gunnery Instructors' School	Upper 27% vs. lower 27% of the candidates	"Of the 60 items, 41 discriminated between the high and the low groups on either the Personal Inventory or the Cornell Selectee Index, with at least a discrimination index (Chi-square) which would have been expected to occur by chance less than 5 times in 100."

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		N R C Neurotic Inventory	
Satter (45)	1400 enlisted Naval personnel	Ratings of the men by their submarine officers	Correlations between inventory scores and the criterion ratings were low and non-significant.
		Humm-Wadsworth Inventory	
Cerf (8)	202 pilots in primary training	Pass vs. fail	Biserial validity coefficients on the 8 Scales ranged from −.21 to .13, only two of them being significant—for the Hysteroid and the Epileptoid Scales.
Cerf (8)	200 pilots in primary training	Pass vs. fail	Biserial validity coefficients on the 8 Scales ranged from −.22 to .16; only those for the Hysteroid and Epileptoid Scales were significant.
Cerf (8)	200 AAF pilots in primary training	Pass vs. fail	Non-significant biserial validity coefficients ranging from −.01 to .16.
Cerf (8)	193 AAF pilots in training	Pass vs. fail*	Non-significant Chi-squares of 1.05 and 4.82 for 4 degrees of freedom. "More than 90% of chance deviations would have been as great."
Cerf (8)	195 AAF pilots in training	Pass vs. fail†	Non-significant Chi-squares of .11 and .34 for 2 d.f. "More than 60% of chance deviations would have been as great .... No significant relationships were found between pilot success in primary flying school and ratings either of Dr. Humm's analyses of temperament integration or of his case summaries."

* Case summaries made by Dr. Humm from the responses to the Inventory were compared with training-course success or failure.

† Case summaries concerning temperamental integration, made by Dr. Humm from the responses to the Inventory, were compared with training-course success or failure.

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		Information Blank	
Cerf (**8**)	275 AAF bombardiers, navigators, and pilots in advanced training	Pass vs. fail	Non-significant biserial validity coefficients of −.05, .07, and .05 were obtained between the test scores and the criterion.
Cerf (**8**)	83 pilots in primary training	Pass vs. fail	Non-significant biserial validity coefficients ranging from .16 to −.21 were obtained. "It does not predict graduation or elimination from primary training."
		Self Descriptive Inventory	
Miles & others (**34**)	3104 Marine trainees	Pass vs. fail	At a cutting score of 6, 80% positives were detected at the expense of 24% false positives and 20% false negatives.
		Confidential Questionnaire	
Williams, Leavitt, & Blair (**72**); also Leavitt, Williams, & Blair (**28**)	666 Marine officer candidates	Pass vs. fail	"The scores . . . correlated with the Platoon Commanders School pass-fail criterion to the extent of .21 (biserial). . . . A small but substantial difference appears between the mean scores of the 395 successful candidates and the 271 unsuccessful candidates, which are 14.6 and 13.8 respectively."
		Personal Preference Questionnaire	
Williams, Leavitt, & Mendola (**74**); also Williams and Leavitt (**69**)	649 Marine Corps officer candidates	Pass vs. fail	Validity coefficients of .31 and .25 were found between the criterion and the Modesty-Egoism and Social Judgment Scales. Critical ratios of the mean score group-differences were 6.5 and 5.1 respectively.
		Minnesota Personality Scale	
Cerf (**8**)	338 AAF pilots in primary training	Pass vs. fail	Non-significant biserial validity coefficients ranged from −.09 to .09. "It appears that the Minnesota Personality Scale, CE 438A, has no value for predicting success in primary pilot training."

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		Minnesota Multiphasic Inventory	
Cerf (8)	400 AAF pilots in primary training	Pass vs. fail	Item validities showed 72 significant *phi*'s out of 699 items. "In view of the apparently unimodal distribution of *phi*'s with a central tendency at zero, it is probable that there are few, if any, genuinely valid items in this collection for the prediction of primary pilot training success."
Jensen & Rotter (22)	1548 Army officer candidates and officers	Officer candidates vs. outstanding officers	Neither the Psychasthenia Scale of the MMPI nor the *C*-Inventory differentiated significantly between the groups of officer candidates and outstanding officers.
		Bell-MMPI-Army Adjustment Inventory	
Altus (1)	3614 Army Special Training Center candidates	Pass vs. fail	The inventory validly distinguished passing from failing ASTC candidates in certain instances. Tetrachoric correlation validity coefficients ran as high as .64.
Altus & Bell (2, 3)	200 Army Special Training Center candidates	Pass vs. fail	The inventory significantly differentiated the passing from the failing candidates, the critical ratio of the mean score group-difference being 8.49. A biserial correlation of .45 between test scores and criterion was also obtained.
		Bernreuter Personality Inventory	
Cerf (8)	800 AAF pilots in primary training	Pass vs. fail	On item analyses, only 6 *phi's* out of 108 reached or exceeded the 5% level of significance. "This instrument contains an insufficient number of valid items for the prediction of primary pilot success to make further scoring measures or statistical analysis worth while."

TABLE II (*continued*)

Source	*Group Tested*	*Performance-Measure*	*Result*
		Personal Audit	
Cerf (8)	271 pilots in primary training	Pass vs. fail	"The biserial coefficients range from −.12 to .09, which are well within the range to be expected of a chance distribution of biserial correlations, the true mean of which is zero . . . . This test is of no value in predicting pilot performance in primary training."
		Inventory of Factors GAMIN	
Cerf (8)	782 AAF pilots in primary training	Pass vs. fail	Non-significant biserial validity coefficients ranged from −.05 to .04. Item validities showed 45 significant *phi*'s out of 424 items. "The Inventory of Factors holds practically no promise as an instrument for predicting graduation-elimination from primary pilot training."
		Inventory of Factors STDCR	
Cerf (8)	1106 AAF pilots in primary training	Pass vs. fail	Biserial validity coefficients ranged from .03 to −.09; 42 out of 304 items were found to have significant *phi*'s. "It appears that the Inventory of Factors STDCR is not promising for predicting graduation-elimination from primary pilot training."
		Guilford-Martin Personnel Inventory	
Cerf (8); also Guilford (19)	945 AAF pilots in primary training	Pass vs. fail	Significant biserial validity coefficients from .10 to .14 were found. Item validities showed 50 significant *phi*'s out of 275 items. "It appears that the Guilford-Martin Personnel Inventory has some promise for predicting graduation-elimination from primary pilot training."

of the distribution of emotional adjustment was more or less eliminated; and this would tend to reduce or conceal such relation between inventory-scores and performance-measures as might actually exist in unselected samples. This would be especially true if the relation between emotional adjustment and performance-measures is curvilinear (i.e., stronger at the lower end than in the middle and upper sections).

Thus, in Graham, Mote, and Berry's (18) use of the Personal Inventory with submarine school recruits, it was noted that about eight per cent of the candidates had, for psychiatric reasons, been refused admission to the school; so that the group which remained to pass or fail in training was presumably "normal." Since *both* the passing and the failing candidates had already been neuropsychiatrically screened, the absence of a significant difference in mean inventory-scores between the two groups may merely testify to the efficiency of the previous screening.

Similarly, Jensen and Rotter (22) used the Minnesota Multiphasic Inventory to try to distinguish officer candidates from outstanding officers, and found no significant difference. But since officer candidates go through several rigorous screenings before they ever reach officer candidate schools, there is little reason to believe that the MMPI could distinguish between the two groups employed in this study.

Again: Satter (45) found that the Personal Inventory did not significantly distinguish between submarine men who received high and low performance ratings by their officers. As Satter points out, "it is conceivable that, since the men were in part selected on the basis of Personal Inventory scores, the low validity coefficients are the result of a curtailed range of adjustment" (45, p. 10).

It may also be pointed out that if the passing and the failing group —or the superior and the less-superior group—are *both* under motivation (conscious or unconscious) to obtain favorable inventory scores, then significant inventory-differences between the two groups are less likely to be found. Thus, as mentioned, Jensen and Rotter (22) found no significant difference in inventory scores of officer candidates vs. outstanding officers. Similarly, in a sample of submarine men (all volunteers, and ordinarily preferring to remain in the submarine branch), Satter (45) found no significant relation between inventory scores and performance-ratings.

2. *Unreliability and invalidity of performance-measures.* Data are lacking by which the reliability and validity of military performance-measures (typically pass-fail in a training-course) could be compared with the reliability and validity of the psychiatric criterion; and in any event, it is too easy to blame the shortcomings of a test on the variable against which it is being correlated. Nevertheless, it may be worth calling attention to the fact that serious unreliability or invalidity of a performance-measure will tend to reduce or conceal whatever correlation may exist between the performance-measure and the inventory scores. In Williams, Leavitt, and Blair's (71) study, for example, scores on the Personal Inventory were found to be negligibly related to ratings of Marine officer candidates on combat proficiency. But such ratings (it seems

fairly well agreed) are themselves of dubious reliability and validity; and it is conceivable that marked improvement of the ratings might raise the negligible relation to a statistically significant and practically useful level (**15**).

3. *Residual relationship.* Individual differences in performance-measures ordinarily depend more largely on differences in aptitude and previous training, than on differences in emotional adjustment.[12] At best, then, it is only what may be termed a residual portion of performance that is dependent on the adequacy of emotional adjustment. In short, a high correlation between personality-inventory scores and performance-measures is not to be expected. By contrast, the psychiatric criterion provides a direct or alternative measure of what the military personality-inventories attempt to measure.

4. *Shift of original criterion.* Finally, it should be remembered that the personality inventories were originally validated against a psychiatric criterion, and not against performance-measures. It is possible that item analysis, making use of performance measures as criteria, could raise the residual relationship between personality-inventory scores and performance-measures to a practically useful level.

Summary and Conclusions

Military applications of personality inventories have yielded enough favorable results to command attention. In contrast, personality inventories in civilian practice have generally proved disappointing. To throw some light on this contrast, the writers undertook to review the available papers on military validation of personality questionnaires. Table I presents a summary of studies making use of a psychiatric criterion (prognosis or diagnosis of neuropsychiatric unfitness for military duty); Table II presents a summary of studies making use of a performance-criterion (most commonly, success in training-courses). Detailed consideration of the various studies leads to the conclusion that both spurious and legitimate factors account for the superior showing of the personality inventories in military practice. With regard to studies making use of a psychiatric criterion, the following factors appear to have played a part in the results obtained:

1. *Criterion contamination* (knowledge of inventory scores at time of making psychiatric prognosis or diagnosis).
2. *Criterion overlap* (duplication of questions asked in the inventory and by the psychiatrist).

[12] There are exceptions to this general statement, especially when the performance is of such a nature as to depend chiefly on persistence of attention, on conscientiousness of effort, or on courage and daring (**44**). But the statement appears to be true at least for the types of performance-measures usually employed in the military studies of personality inventories.

3. *Use of extreme or atypical groups.* The use of extreme groups or of groups containing an atypically large proportion of psychiatrically "positive" cases leads to atypically favorable results (lower incidence of "false positives," higher biserial or tetrachoric correlation between inventory and criterion).

4. *Differential motivation.* In some studies making use of two contrasted groups, the motivation of the "abnormal" group toward self-incriminating responses, and/or of the normal group toward self-inflationary responses, probably led to exaggerated differences between the groups and a spuriously high index of validity for the personality inventory.

5. *Statistical inadequacies.* In some of the studies, questionable or inflated indexes of validity were obtained through: calculation of biserial correlations for samples containing an extraordinarily high proportion of abnormals; use of samples containing very small numbers of cases in the diagnostic categories; reliance upon the critical ratio or *t*, without considering the degree of relation connoted by the given *C.R.* or *t;* and use of the original standardization sample (instead of a fresh sample) for validation purposes.

6. *Lenient evaluation of false-positive results.* Except in samples containing an unusually large proportion of psychiatrically "positive" cases, the number of cases *falsely* classified as positive by the inventory generally exceeds, by a great deal, the number *correctly* classified as positive. Each false-positive case may, however, be viewed as a misclassified normal case; the percentage of misclassification (percentage of false-positive to total normal) is generally small. In most of the studies, the percentage-interpretation prevails, rather than the more practical interpretation in terms of number of cases.

7. *Neglect of "false-negative" cases.* In most validity studies (both military and civilian), the number and percentage of false-negative cases are not recorded. Since false negatives are a more serious liability in the military situation than in the civilian, the neglect of false negatives may lead to undue optimism in evaluating the military studies.

8. *Sample heterogeneity.* As compared with civilian examinees, the armed forces' unscreened recruits frequently included a larger proportion of lower-end (and comparatively easily diagnosable) cases, such as unemployables, "bums," alcoholics, frank neurotics, etc.

9. *Lower level of intelligence.* It appears likely that the comparatively direct questions of personality inventories have greater validity when used with less intelligent or more naïve individuals such as are found in military samples, rather than the selected samples commonly used in civilian studies.

10. *Honesty of response.* It appears likely that members of the armed services answered the personality inventories with less distortion than civilians ordinarily do. Among possible reasons for this is the fact that the direct penalties for falsification are greater in the military situation than in the civilian.

11. *Specialized design, validation, and standardization.* The inventories in most common use were specifically designed for military groups, and often were modified when applied to a group different from the standardization-group. Moreover, external criterion groups were used for validation purposes, instead of merely the criterion of "internal consistency."

12. *Realistic application.* In general, the inventories were applied for screening only, and not for elaborate personality analysis.

While the military investigators should be credited with the favor-

able factors mentioned above, it would be unfair to charge them with ignorance or neglect of the other factors. The exigencies of military circumstances generally prevented the application of any completely ideal methodology, even when the experimenter was keenly aware of the shortcomings of his samples or his procedures.

In contrast to their success in relation to the psychiatric criterion, the military personality inventories proved generally ineffective for predicting performance-measures (such as successful completion of a training-course). The reasons for this difference appear to be as follows:

1. *Prior elimination of abnormals.* Most men selected for training-courses had already been screened or otherwise certified as to minimum necessary emotional fitness.
2. *Unreliability or invalidity of the performance-measures.* This may be a contributing factor to the unfavorable results, at least in a few instances.
3. *Residual relationship.* Individual differences in performance-measures ordinarily depend more largely on differences in aptitude and previous training, than on differences in emotional adjustment. The relationship between personality-inventory scores and performance-measures, in consequence, is at best of only residual strength.
4. *Shift of original criterion.* The original validation of the personality inventories was in terms of a psychiatric criterion, and not in terms of performance measures.

In conclusion it appears that, while the experimental or statistical shortcomings of many of the military studies justify a cautious attitude toward the results obtained, the fact cannot be ignored that the inventories usually did make some definite contribution to psychiatric screening. The success of the inventories in the military situation encourages the hope that similar inventories may prove equally useful in civilian practice. Military experience suggests emphasis on the following points:

1. Personality questionnaires should be especially designed for the group to whom they are applied, and should be validated against dependable external criteria. Criterion-contamination should be guarded against; and criterion-overlap, if it occurs, should be taken into account in evaluating the findings.
2. Special attention should be given to persuading or inducing respondents to answer the inventory-items as truthfully as they can.
3. Personality inventories may possibly be more effective when used with relatively uneducated and less intelligent groups, than with groups that are more sophisticated.
4. The users of personality inventories should realize that only limited and specialized demands may be made on the inventory technique; and that broad and incisive personality diagnosis is still the specialty of the trained clinician employing subtler and more comprehensive psychological techniques.

BIBLIOGRAPHY

1. ALTUS, W. The adjustment of Army illiterates. *Psychol. Bull.*, 1945, **42**, 461–476.
2. ALTUS, W. D., & BELL, H. M. The validity of certain measures of maladjustment in an Army Special Training Center. *Psychol. Bull.*, 1945, **42**, 98–103.
3. ALTUS, W. D., & BELL, H. M. An analysis of four orally administered measures of adjustment. *Educ. psychol. Msmt.*, 1947, **7**, 101–115.
4. BENTON, A. L. The Minnesota Multiphasic Inventory in clinical practice. *J. nerv. ment. Dis.*, 1945, **102**, 416–420.
5. BENTON, A. L., & PROBST, KATHRYN A. A comparison of psychiatric ratings with Minnesota Multiphasic Inventory scores. *J. abnorm. soc. Psychol.*, 1946, **41**, 75–78.
6. BERRY, R. N., LEAVITT, H. J., & MOTE, F. A. *The comparability of Formats A and B of the Personal Inventory.* OSRD Report No. 3582. Project Report No. 6. Providence, R.I.: Brown University, April 21, 1944.
7. BOBBITT, J. M., & NEWMAN, S. H. Comparative hospital records of two groups differentiated by psychological tests. *J. consult. Psychol.*, 1947, **11**, 292–298.
8. CERF, A. Z. Personality inventories. In J. P. Guilford (Ed.), *Printed classification tests.* Army Air Forces Aviation Psychology Program Research Reports, No. 5. Washington: 1947. Pp. 577–621.
9. CONRAD, H. S. Overview. *Rev. educ. Res.*, 1947, **17**, 1–3.
10. COVILLE, W. J. A study of the effectiveness of the Maritime Service Inventory as a screening device at the U. S. Maritime Service Training Station, St. Petersburg, Florida. In G. Killinger (Ed.), *The psychobiology of the War Shipping Administration.* Appl. Psychol. Monogr., No. 12. Stanford University: Stanford University Press, 1947. Pp. 101–114.
11. DYNES, J. B. Mental breaking point. *New Eng. J. Med.*, 1946, **234**, 42–45.
12. ELLIS, A. The validity of personality questionnaires. *Psychol. Bull.*, 1946, **43**, 385–440.
13. ELLIS, A. Personality questionnaires. *Rev. educ. Res.*, 1947, **17**, 53–63.
14. FISKE, D. W. Validation of Naval Aviation Cadet Selection Tests against training criteria. *J. appl. Psychol.*, 1947, **31**, 601–616.
15. FLANAGAN, J. C. Summary of discussion in the session on instruments of measurement. In G. Kelley (Ed.), *New methods in applied psychology.* College Park, Md.: University of Maryland, 1947. Pp. 190–194.
16. GILLILAND, A. R. Problems of personality. *J. abnorm. soc. Psychol.*, **23**, 369–378.
17. GOUGH, H. G. Diagnostic patterns in the Minnesota Multiphasic Inventory. *J. clin. Psychol.*, 1946, **2**, 23 37.
18. GRAHAM, C. H., MOTE, F. A., & BERRY, R. N. *The relation of selection test scores to tank escape performance: submarine school.* OSRD Report No. 3262. Providence, R. I.: Brown University, Jan. 31, 1944.
19. GUILFORD, J. P. Some lessons from aviation psychology. *Amer. Psychologist*, 1947, **3**, 3–11.
20. HARRIS, H. J. The Cornell Selectee Index—an aid in psychiatric diagnosis. *Annals. N. Y. Acad. Sci.*, 1946, **46**, 594–603.
21. HEATHERS, G. L. Personality and adjustment studies. In S. W. Bijou (Ed.), *The psychology program in Army Air Forces convalescent hospitals.* Army Air Forces Aviation Psychology Program Research Reports, No. 15. Washington: 1947. Pp. 59–85.
22. JENSEN, M. B., & ROTTER, J. B. The value of thirteen psychological tests

in officer candidate screening. *J. appl. Psychol.*, 1947, **31**, 312–322.

23. KILLINGER, G. G. (Ed.) *The psychobiological program of the War Shipping Administration.* Appl. Psychol. Monogr., No. 12. Stanford University: Stanford University Press, 1947.
24. KILLINGER, G., & ZUBIN, J. The psychobiological program of the War Shipping Administration. In G. Killinger (Ed.), *The psychobiological program of the War Shipping Administration.* Appl. Psychol. Monogr., No. 12. Stanford University, Calif.: Stanford University Press, 1947, Pp. 23–32.
25. KORNHAUSER, A. Replies of psychologists to a short questionnaire on mental test developments, personality inventories, and the Rorschach test. *Educ. psychol. Msmt.*, 1945, **5**, 3–15.
26. LEAVITT, H. C. A comparison between the neurospsychiatric screening adjunct (NSA) and the Cornell Selectee Index (Form N). *Amer. J. Psychiat.*, 1946, **103**, 353–357.
27. LEAVITT, H. J., & ADLER, N. *Validation of officer selection tests by means of combat proficiency ratings. Progress Report No. 2: Final analysis.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-620 (Sub. No. 135), May 16, 1946.
28. LEAVITT, H. J., WILLIAMS, S. B., & BLAIR, C. L. *Application of the NDRC Personal Inventory and other measures in the selection of Officer Candidates. Progress Report No. 5: The use of school grades as a criterion of test validity.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-386 (Sub. No. 75), January 15, 1946.
29. LEAVITT, H. J., WILLIAMS, S. B., & LIPKIN, N. J. *Application of the NDRC Personality Inventory to Marine Corps Officer Candidates. Progress Report No. 3: Data on Final 671 Candidates.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-386 (Sub. No. 75), Dec. 20, 1945.
30. LEPLEY, W. M. *Psychological research in the theatres of War.* Army Air Forces Aviation Psychology Research Reports, No. 17. Washington 1947.
31. LEVERENZ, C. W. Minnesota Multiphasic Personality Inventory—an evaluation of its usefulness in the psychiatric service at a station hospital. *War. Med.*, 1943, **4**, 618–629.
32. MALLER, J. B. Personality tests. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald Press, 1944.
33. MANSON, M. P., & GRAYSON, H. M. The "sick book rider" in an oversea military prison. *Psychosom. Med.*, 1946, **8**, 414–416.
34. MILES, D. W., *et al.* The efficiency of a high-speed screening procedure in detecting the neuropsychiatrically unfit at a U. S. Marine Corps recruit training depot. *J. Psychol.*, 1946, **21**, 243–268.
35. MITTELMANN, B. The Cornell Selectee Index: short form to be used at induction, at reception and during hospitalization. *War Psychiatry: Proceedings of the second brief Psychotherapy Council.* Chicago: Institute for Psychoanalysis, 1944.
36. MODLIN, H. C. A study of the Minnesota Multiphasic Inventory in clinical practice. *Amer. J. Psychiat.*, 1943, **103**, 748–769.
37. MORRIS, W. W. A preliminary evaluation of the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1947, **3**, 370–374.
38. MOTE, F. A., BERRY, R. N., & GRAHAM C. H. *Results obtained from testing recruits with the New London-NDRC Questionnaire at the Newport Naval Training Station.* OSRD Re-

port No. 3040. Providence, R. I.: Brown University, December 6, 1943.

39. OWENS, W. A. Item form and "false-positive" response on a neurotic inventory. *J. clin. Psychol.*, 1947, **3**, 264–269.
40. OWENS, W. A., & ZIRKLE, G. A. The form of items and distribution of false positive scores on a neurotic inventory. In G. Kelly (Ed.), *New methods in applied psychology.* College Park, Md.: University of Maryland, 1947. Pp. 190–194.
41. PAGE, H. E. Detecting psychoneurotic tendencies in Army personnel. *Psychol. Bull.*, 1945, **42**, 645–658.
42. ROBACK, A. A. Personality tests—whither? *Character & Pers.*, 1933, **1**, 214–224.
43. ROSENZWEIG, S. A basis for the improvement of personality tests, with special reference to the Masculinity-Femininity battery. *J. abnorm. soc. Psychol.*, 1932, **26**, 415–421.
44. SATTER, G. A. *An evaluation of the Personal Inventory for predicting success in parachute school.* OSRD Report No. 4870. Princeton, N. J.: Research & Statistical Laboratory, College Entrance Examination Board, March 28, 1945.
45. SATTER, G. A. *An evaluation of the Personal Inventory and certain other measures in the prediction of Submarine Officer's evaluations of enlisted men.* OSRD Report No. 5557. Princeton, N. J.: Research & Statistical Laboratory, College Entrance Examination Board, September 7, 1945.
46. SCHMIDT, H. O. Test profiles as a diagnostic aid: the Minnesota Multiphasic Inventory. *J. appl. Psychol.*, 1945, **29**, 115–131.
47. SHAFFER, L. F. Psychological studies of anxiety reaction to combat. In F. Wickert (Ed.), *Psychological research on problems of redistribution.* Army Air Forces Aviation Psychology Program Research Reports, No. 14. Washington: 1947. Pp. 93–131.
48. SHIPLEY, W. C., & GRAHAM, C. H. *Final report in summary of research on the ...rsonal Inventory and other tests.* OSRD Report No. 3963. Project Report No. 10. Providence, R. I.: Brown University, August 1, 1944.
49. SHIPLEY, W. C., GRAY, F. E., & NEWBERT, N. *Item analysis and evaluation of the scoring stencil of the Personal Inventory.* OSRD Report No. 3315. Project Report No. 4. Providence, R. I.: Brown University, Feb. 14, 1944.
50. SHIPLEY, W. C., GRAY, FLORENCE E., & NEWBERT, NANCY. *The Personal Inventory, Short Form (Format C): derivation and preliminary psychiatric validation.* OSRD Report No. 3390. Providence, R. I.: Brown University, March 15, 1944.
51. SHIPLEY, W. C., GRAY, FLORENCE E., & NEWBERT, NANCY. *A comparison of Personal Inventory scores with service records one year after testing.* OSRD Report No. 3755. Project Report No. 8. Providence, R. I.: Brown University, June 10, 1944.
52. SHIPLEY, W. C., *et al.* The Personal Inventory. *J. clin. Psychol.*, 1946, **2**, 318–322.
53. SMITH, K. R., & VOSS, H. A. *Item differentiation and derivation of a 50-item scoring key for the Officer Personal Inventory, Form 1.* Project N-117, Applied Psychology Panel, NDRC. Memorandum No. 4, February 8, 1945.
54. STOLUROW, L. M., IRION, A. L., & PASCAL, G. R. The selection and training of gunnery instructors. In N. Hobbs (Ed.), *Psychological research on flexible gunnery training.* Army Air Forces Aviation Psychology Program Research Reports, No. 10. Washington: 1947. Pp. 337–381.

55. STOLUROW, L. M., & SCHRADER, W. B. The selection of gunners. In N. Hobbs (Ed.), *Psychological research on flexible gunnery training.* Army Air Forces Aviation Psychology Program Research Reports, No. 10. Washington: 1947. Pp. 61–98.

56. STONE, L. J., & MALAMENT, M. The construction of the Maritime Service Inventory. In G. Killinger, *The psychobiological program of the War Shipping Administration.* Appl. Psychol. Monogr., No. 12. Stanford University, Calif.: Stanford University Press, 1947. Pp. 89–100.

57. THORPE, L. P. *Psychological foundations of personality.* New York: McGraw-Hill, 1938.

58. TRAXLER, A. E. *The use of tests and rating devices in the appraisal of personality.* New York: Educational Records Bureau, 1942. Educational Research Bulletin No. 23, Rev. Ed.

59. VARNEY, H. I., & STONE, L. J. The psychobiological program at the U. S. Maritime Service Training Station, Sheepshead Bay, N. Y. In G. Killinger (Ed.), *The psychobiological program of the War Shipping Administration.* Appl. Psychol. Monogr., No. 12. Stanford University, Calif.: Stanford University Press, 1947. Pp. 33–60.

60. VERNON, P. E. The attitudes of the subject in personality testing. *J. Appl. Psychol.*, 1934, **18**, 165–167.

61. WEIDER, A., & WECHSLER, D. The Cornell Indices and the Cornell Word Form: 2. Results. *Annals N. Y. Acad. Sci.*, 1946, **46**, 579–591.

62. WEIDER, A., *et al.* The Cornell Selectee Index. *J. Amer. med. Ass.*, 1944, **124**, 224–228.

63. WEIDER, A., *et al.* Cornell Service Index. *War Med.*, 1945, **7**, 209–213.

64. WEINSTOCK, H. I., & WATSON, R. I. The usefulness of the Cornell Selectee Index at the neuropsychiatric unit of a naval training station. *U. S. Naval Med. Bull.*, 1946, **46**, 1583–1588.

65. WEXLER, M. Measures of personal adjustment. In D. B. Stuit (Ed.), *Personnel research and test development in the Bureau of Naval Personnel.* Princeton, N. J.: Princeton Univ. Press, 1947. Pp. 126–174.

66. WEXLER, M., OWENS, W. A., & PORTER, R. B. Test procedures for the psychiatric screening of naval personnel. In G. Kelly (Ed.), *New methods in applied psychology.* College Park, Md.: University of Maryland, 1947. Pp. 60–66.

67. WILEY, L. N., & TRIMBLE, O. C. The ordinary objective test as a possible criterion of certain personality traits. *Sch. & Soc.*, 1936, **43**, 446–448.

68. WILLIAMS, S. B., & LEAVITT, H. J. *Application of the NDRC Personal Inventory to Marine Corps officer candidates: preliminary validation.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-386 (Sub. No. 75), December 26, 1944.

69. WILLIAMS, S. B., & LEAVITT, H. J. Methods of selecting Marine Corps officer candidates. In G. Kelly (Ed.), *New methods in applied psychology.* College Park, Md.: University of Maryland, 1947. Pp. 96–99.

70. WILLIAMS, S. B., LEAVITT, H. J., & ADLER, N. *Application of the NDRC Personal Inventory and other measures in the selection of officer candidates. Progress Report No. 6: Final report and summary.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-386 (Sub. No. 75), January 19, 1946.

71. WILLIAMS, S. B., LEAVITT, H. J., & BLAIR, C. R. *Validation of officer selection tests by means of combat proficiency ratings. Progress Report No. 1: The prediction of successful combat leadership.* Medical Field Research

Laboratory, Camp Lejeune, N. C. M. & S. Research Project No. X-620 (Sub. No. 135), January 18, 1945.

72. WILLIAMS, S. B., LEAVITT, H. J., & BLAIR, C. L. *The development and validation of the MFRL Confidential Questionnaire.* Medical Field Research Laboratory, Camp Lejeune, N. C., M. & S. Research Project No. X-615 (Sub. No. 133), December 21, 1945.

73. WILLIAMS, S. B., LEAVITT, H. J., & MENDOLA, G. V. *Application of the NDRC Personal Inventory to Marine Corps officer candidates. Progress Report No. 2: Further validation.* Medical Field Research Laboratory, Camp Lejeune, N. C., M. & S. Research Project No. X-386 (Sub. No. 75), November 27, 1945.

74. WILLIAMS, S. B., LEAVITT, H. J., & MENDOLA, G. V. *Application of the Personal Inventory and other measures in the selection of officer candidates. Progress Report No. 4: A follow-up study of the Personal Preferences Technique.* Medical Field Research Laboratory, Camp Lejeune, N. C. M. &. S. Research Project X-386 (Sub. No. 75), January 10, 1946.

75. WOLFF, H. G., *et al. The Selectee Index: a method for quick testing of selectees for the armed forces.* Committee on Medical Research of the Office of Scientific Research and Development, August 31, 1943.

76. ZUBIN, J. Recent advances in screening the emotionally maladjusted. *J. clin. Psychol.*, 1948, 4, 56–62.

THE LATIN SQUARE PRINCIPLE IN THE DESIGN AND ANALYSIS OF PSYCHOLOGICAL EXPERIMENTS

DAVID A. GRANT[1]
University of Wisconsin

INTRODUCTION

The latin square, as such, may be used infrequently in psychological investigation, but the basic principles of experimental design and analysis embodied in the latin square may be used to increase greatly the efficiency of many types of psychological research. This fact has already been pointed out (6, 15), but a more extensive discussion seems desirable. It is the purpose of this paper to outline some of the relevant features of the latin square and to illustrate how these features apply in several kinds of psychological experiments. It will be assumed that the reader has already acquired a familiarity with the rudiments of analysis of variance (3, 11, 14).

A latin square is an arrangement of latin letters in rows and columns such that each letter appears once and only once in each row and each column (4, 7). As an example, a five-by-five latin square is given below.

C	*E*	*B*	*D*	*A*
E	*A*	*C*	*B*	*D*
B	*D*	*E*	*A*	*C*
A	*B*	*D*	*C*	*E*
D	*C*	*A*	*E*	*B*

Such latin squares are typically used in agricultural field experiments to control soil variability. Thus A, B, C, D and E typically represent the independent variable, say of five different fertilizer treatments used on 25 small plots of ground arranged as in the above square. Any natural soil gradients parallel to the rows or the columns of the square will introduce irrelevant variability into the plot yields, but variation from such gradients can be eliminated statistically so that the precision of the comparison between treatments will be increased. The analysis of variance for this latin square will be presented later.

Some of the important features of the latin square stand out when this design is contrasted with a three-factor experiment in which each level of each factor appears with each level of every other factor (4, 7). If rows, columns, and treatments are considered as independent vari-

[1] This paper was completed during a research leave supported by the Graduate Research Committee of the University of Wisconsin from special funds provided by the State Legislature for 1947–48.

ables or "main effects," both designs contain three independent variables which are manipulated by the experimenter. The three independent variables are, moreover, orthogonal to or statistically independent of each other in both designs. The crucial difference between the two designs appears in the evaluation of the interactions. In the factorial design, all three first-order interaction effects and the second-order interaction effect can be separated and evaluated.[2] The interactions are all orthogonal to the main effects and to each other. This is not the case with the latin square. In the latin square the interactions, if present,[3] are confounded (or mixed in) both with the effects of the single independent variables and also with each other. The interactions ordinarily cannot be evaluated in a latin square design. Although the confounding effects balance out in complete sets of squares, for any given square, the confounding may be serious enough to counteract or to enhance a main effect. Although this is relatively well-controlled in agricultural applications, in psychological applications the experimenter must at least be aware of the consequences of interactions confounded with main effects. The exact nature of this confounding will be demonstrated in connection with the 2×2 latin square.

The 2×2 Latin Square

The 2×2 latin square gives a striking illustration of the confounding principle just mentioned. There are two 2×2 latin squares, one of which is shown below:

	I	*II*
1	A_{11}	B_{12}
2	B_{21}	A_{22}

The subscripts indicate the row and column position of each entry in the square. Analysis of this square is very instructive. Suppose that *A* and *B* represent two fertilizer treatments. Two parallel analyses of this square are given in Table I, considering it first as a factorial design and then as a latin square design. The correction term,

$$C = \tfrac{1}{4}(A_{11} + B_{12} + B_{21} + A_{22})^2.$$

Examination of Table I reveals that the row and column sums of squares are identical, but that the term known as "interaction" in the factorial design is the same as that known as the "treatment effect" in

[2] Assuming an error estimate from replication.

[3] In agricultural work such interaction might arise from a fertility gradient along a diagonal of the square.

TABLE I

OUTLINED ANALYSIS OF 2×2 LATIN SQUARE CONSIDERED (1) AS A FACTORIAL DESIGN AND (2) AS A LATIN SQUARE DESIGN

Source of Variation	(1) *Factorial Design* *Sum of Squares*	(2) *Latin Square Design* *Sum of Squares*
Rows	$\frac{1}{2}(A_{11}+B_{12})^2+\frac{1}{2}(B_{21}+A_{22})^2-C$	$\frac{1}{2}(A_{11}+B_{12})^2+\frac{1}{2}(B_{21}+A_{22})^2-C$
Columns	$\frac{1}{2}(A_{11}+B_{21})^2+\frac{1}{2}(B_{12}+A_{22})^2-C$	$\frac{1}{2}(A_{11}+B_{21})^2+\frac{1}{2}(B_{12}+A_{22})^2-C$
Row × Column Interaction	$\frac{1}{2}(A_{11}+A_{22})^2+\frac{1}{2}(B_{12}+B_{21})^2-C$	
Treatment		$\frac{1}{2}(A_{11}+A_{22})^2+\frac{1}{2}(B_{12}+B_{21})^2-C$
Total	$A_{11}^2+B_{12}^2+B_{21}^2+A_{22}^2-C$	$A_{11}^2+B_{12}^2+B_{21}^2+A_{22}^2-C$

the latin square. This is always the case with the 2×2 latin square; the treatment effect is identical with the row × column interaction. This identity represents total confounding of these two sources of variation. It is easy to see, furthermore, that the interaction of any pair of independent variables is completely confounded with the third independent variable. Obviously one could not determine the row × column interaction effect because of the complete confounding of that with the treatment effect. Probably more important, one could not determine treatment effects because of the possible presence of row × column interaction.

In the larger squares confounding occurs, but it is not complete. The agricultural experimenter is not greatly inconvenienced by this type of confounding, because he takes pains to use larger randomly selected squares and avoids the use of systematic squares (**4**, sec. 34) which emphasize this effect.[4] Moreover he typically uses both rows and columns to effect a double elimination of a single irrelevant variable, soil heterogeneity. The psychological experimenter who uses latin squares to control two *different* irrelevant variables may find, however, that confounding can cause him serious difficulties. This is particularly true if he uses the 2×2 latin square.

The 2×2 latin square arises rather frequently in experimental psychology. Often each *S* is run through an experimental (*E*) and a control (*C*) procedures. The careful investigator will usually run his half of his *Ss* in the sequence *E*–*C*, and half in the sequence *C*–*E*, and

[4] The selection of squares at random reduces the expected value of confounding to zero. In the case of a $N \times N$ latin square, the expected value of the treatment mean square equals the error variance plus N times the treatment variance. The technique of selecting a square at random is described by Fisher and Yates (**5**, p. 13).

the 2×2 latin square[5] results. If the *sequence* (rows) in which the two procedures are run interacts with *ordinal position effects* (columns)—involving perhaps habituation, practice, fatigue, etc.—this interaction will be *completely confounded* with the difference between the control and experimental procedures. This follows from the fact noted above, that in the two by two latin square the interaction of any two of the three factors, rows, columns, and letters, is completely confounded with the third factor. It should be remarked that the confounding is a property of the design and will be present no matter whether the data are completely analyzed by means of analysis of variance or are incompletely analyzed by means of critical ratios or *t*-tests.

A recent study on the T.A.T. (9) serves as an example in which such effects may be present. The intention was to discover whether there were differences in the themas evoked by the cards numbered 1–10 as compared with the themas evoked by the cards numbered 11–20.

Thus, the differential stimulus value of the cards numbered 1–10 and cards 11–20 was the independent variable of the experiment. Half of the *Ss* were given the cards in the normal sequence 1–10 then 11–20, and half were given the cards in reverse sequence 11–20 then 1–10 because the actual *order* in which the cards were presented was irrelevant to the experiment. This can be presented diagramatically as follows:

		Session	
		I	*II*
Sequence	*Normal*	1–10	11–20
	Reverse	11–20	1–10

The first session is indicated by *I*, and the second session by *II*. Briefly stated, the results showed that more "personal," hostile, and insecure themas appeared in the second ten, and this difference held especially during the second session.

The presence of row×column or session×sequence interaction in this experiment is not known, but there may be grounds for suspecting its existence. For example, *Ss* might have been generally cautious and unsure of themselves in the first session which could have depressed the frequency of personal themas in both sets of cards. In the second session, *Ss*, in general, might have been more relaxed and more willing to deliver anti-social themas especially if they were stimulated with the complex cards 11–20. If the cards were presented in the reversed order, however, the more commonplace cards of the first ten might appear dull

[5] There is only one 2×2 latin square in this context.

and conventional after the *S* had faced the bizarre second ten, and, in consequence, he might have tended to give more conventional stereotyped themas. If any such process takes place, it may constitute a row-column interaction effect which could overemphasize any differences in the stimulus value of the first and second ten cards of the T.A.T.

The type of interaction described above can have most serious consequences. Opportunities for such interactions to arise are frequent in psychological experimentation. Whether serious interactions are commonly obtained is hard to say. The experimenter who uses two procedures in the two possible presentation sequences must, however, be alert to consider such an eventuality in interpreting his experimental data.

Individual Differences and the Latin Square

A very useful application of the latin square to psychological research arises in connection with the simultaneous control of individual differences and temporal order of presentation of procedures. Just as the agricultural researcher faced with the constant problems of soil heterogeneity has turned to the latin square, so the psychologist faced with his perpetual problem of individual differences, can also control a difficult extraneous variable in his experiments by use of the same device. Actually, as Garrett and Zubin (6, p. 242) have pointed out, balanced orders of presentation, permuted double-fatigue orders, the classical *ABBA* order, etc. have long been used by careful experimenters, and these all embody certain latin square principles. Furthermore, a number of excellent papers (summarized in 12) have shown how repeated scores and matched groups offer opportunities for reducing the size of error estimates to increase the efficiency of experiments. This paper presents an extension of these principles well-known to the researcher.

First, as an artificially simple example, using the 5×5 latin square shown earlier, suppose that the dependent variable is the amount of aggressive behavior observed in nursery school children in five controlled social situations. The five social situations constitute the independent variable. Because the children may be expected to show consistent individual differences in aggressiveness, a factor irrelevant to the experiment, it is desirable that each child serve as his own control or go through each of the five different situations. Furthermore, let us suppose that there may be some practice or adaptation effect in the situations. This would rule out the procedure of going through the five situations in one constant sequence or order. Let the social situations be designated *A*, *B*, *C*, *D*, and *E*. If five children are available, each child

could be assigned to a single row of the latin square above, and run through the situations in the order indicated below:

		Order				
		I	*II*	*III*	*IV*	*V*
Child	*1*	*C*	*E*	*B*	*D*	*A*
	2	*E*	*A*	*C*	*B*	*D*
	3	*B*	*D*	*E*	*A*	*C*
	4	*A*	*B*	*D*	*C*	*E*
	5	*D*	*C*	*A*	*E*	*B*

His aggressive behavior score would then be entered into a table set up in exactly the same pattern. In the above arrangement, ***individual differences*** would then produce inter-row variation with four degrees of freedom, ***practice effect*** would produce inter-column variation with four degrees of freedom, and the experimental factor, ***social situation***, latin letters, with four degrees of freedom would be orthogonal to each of these. The error variation would be estimated with twelve degrees of freedom.

Computation of the analysis of variance is straightforward. Let ΣX^2 be the sum of all squared aggression scores. The correction factor, $C = 1/25\ (\Sigma X)^2$. Then, letting the totals for *Children* have arabic subscripts, the totals for *Order* have latin numeric subscripts, and the totals for *Situation* have latin literal subscripts, the sums of squares are as follows:

1. $SS_{Tot} = \sum X^2 - C.$
2. $SS_{Children} = 1/5(T_1^2 + T_2^2 + \cdots + T_5^2) - C.$
3. $SS_{Order} = 1/5(T_I^2 + T_{II}^2 + \cdots + T_V^2) - C.$
4. $SS_{Situations} = 1/5(T_A^2 + T_B^2 + \cdots + T_E^2) - C.$
5. $SS_{Error} = SS_{Tot} - (SS_{Children} + SS_{Order} + SS_{Situations}).$

Row × column interaction or individual differences in practice effect, if present, cannot be assessed. It may serve to enlarge or to shrink the error estimate and the mean square for the experimental factor. The other interactions are likewise confounded so that if certain situations favor some children but not all, or if certain situations are reacted to differently early in the sequence from the way they are reacted to later in the sequence, these interactions cannot be tested. If present, they may influence the size of the error term in the analysis of variance.

Complete Sets of Squares

The experiment with nursery school children was a direct application of the latin square. More commonly the latin square will not be used directly, but the latin square principle will prove useful. As an example, consider the following investigation **(8)**. It was desired to study the process of learning to learn nonsense syllables. Four experimental

procedures were used. (They involved three variations of group motion picture presentation versus traditional individual sessions with the memory drum.) After the experimental procedures all *Ss* were given the same three successive lists to learn as a test. The number of trials necessary to learn to a criterion of two successive perfect repetitions was the score. (After pre-training the distributions of these scores were not badly skewed.) The three test lists, A, B, and C might be expected to vary in difficulty, and moreover, the *Ss* would still be changing in learning skill while memorizing the three test lists. These two factors were irrelevant to the experiment and should be controlled in order to eliminate the variance they introduced from the error estimate. Attention will first be focussed on one of the four experimental groups. To balance out list difficulty each of the three lists should be used in each of the three positions. A latin square was thus suggested with three *Ss* taking the three lists in three different sequences, but three *Ss* would scarcely be adequate for reliable measures in a rote learning experiment. To overcome this difficulty all six permutations of the three lists —in effect two latin squares—were used. It was therefore necessary to use a number of *Ss* which was a multiple of six; *e.g.*, thirty. Thus the data for this group consisted of 90 scores in a 30 row by 3 column table, where each row represented a different subject, and the three columns represented the first, second and third lists in order of presentation.

Let a score from the above-mentioned table be designated by X_i, the sum of scores for *Ss* by $T_1, T_2, \ldots, T_{30}$, the sum of scores for the three lists by T_A, T_B, and T_C, and the sum of scores for the three ordinal positions in the presentation sequence as T_{I}, T_{II}, and T_{III}. The sums of squares for analysis of variance were found as follows:

$$(1) \qquad SS_{total} = \sum_{i=1}^{90} X_i^2 - C. \qquad 89/\mathrm{df}$$

$$C = \frac{\left(\sum_{i=1}^{90} X_i\right)^2}{90}$$

$$(2) \qquad SS_{subjects} = 1/3 \sum_{i=1}^{30} T_i^2 - C \qquad 29/\mathrm{df}^{6}$$

$$(3) \qquad SS_{lists} = 1/3(T_A^2 + T_B^2 + T_C^2) - C \qquad 2/\mathrm{df}$$

$$(4) \qquad SS_{order} = 1/30(T_{\mathrm{I}}^2 + T_{\mathrm{II}}^2 + T_{\mathrm{III}}^2) - C \qquad 2/\mathrm{df}$$

$$(5) \qquad SS_{error} = SS^2_{total} - (SS_{subjects} + SS_{lists} + SS_{order}) \qquad 56/\mathrm{df}$$

[6] The $SS_{subjects}$ and its 29 df can be broken down into $SS_{sequences}$ with 5 df and $SS_{individuals\ within\ sequences}$ with 24 df, using the procedure described by Kogan (10).

In this case the sums of squares for the three "main effects," subjects, lists, and order would not be inflated by confounded interactions even if appreciable interactions were present. The sum of squares for error did not contain confounded interaction but was restricted to "pure error." This fortunate state of affairs arises because all possible combinations were used an equal number of times. Interactions, however, could not be assessed, except perhaps with considerable loss of data because of partial confounding.

The final analysis in the experiment was essentially a "between groups"—"within groups" analysis (11, 14). Each of the other three experimental groups was treated in essentially the same way, so that four parallel analyses were made. The four group totals were T_α, T_β, T_γ, and T_δ, and the differences between groups was assessed in terms of:

$$(6)\quad SS_{\text{between groups}} = 1/90(T_\alpha^2+T_\beta^2+T_\gamma^2+T_\delta^2) - \frac{(T_\alpha+T_\beta+T_\gamma+T_\delta)^2}{360}$$

with 3 df.

The mean square between groups was divided by the pooled error mean square which was obtained by adding together the four SS_{error} for the four groups and dividing this error total by 4×56 or 224. Where the resulting F was significant the differences between groups are significant. But this test has a rather narrow scope. A second F should be computed, using the *between groups* mean square again as the numerator and a pooled *Ss within groups* mean square as the denominator. (This denominator can be set up by adding the four $SS_{subjects}$ for the four groups and dividing by 4×29 or 116.) If this second F is not significant the differences between groups may be accounted for in terms of reliable differences between *Ss*. If the second F is significant, it means that the differences between groups exceed the differences which might be attributed to consistent variation between the individual *Ss*.

The special virtue of this design was that in addition to balancing out extraneous factors such as list difficulty, and order effects, these and individual differences (subject variance) were removed from the error estimate so that considerable precision was gained. This procedure can often be followed, and when complete sets of permutations are used, all interactions within such sets of squares cancel out so that they are not confounded with the main effects.

Replicated Squares

When it is impossible to use complete sets of permutations in a design similar to that described above there remain two options. On

the one hand several different randomly selected latin squares might be used, or on the other hand, the same square might be used over several times. Each procedure has its merits. Using several different squares will almost insure that the interactions effects of each square will be cancelled out by those of the others. Using one square several times enables the experimenter to obtain an error estimate containing only certain interactions, and the variables may be such that the experimenter may have grounds to suppose that these interactions are of no consequence.

An example of the repeated use of the same latin square appears in a recently completed experiment of Corrigan and Brogden (2). The dependent variable was the precision of linear horizontal pursuit movements which was studied as a function of the independent variable, the angle of the path from the line normal to the body of the *S*. Seven angles were used, and the *Ss* were given tests at these seven angles in seven sequences such that each angle occurred once and only once in each ordinal position of the sequence. A 7×7 latin square was thus formed with *Ss* or sequences for rows, ordinal position within the sequence of columns and angle of pursuit for latin letters. Instead of having one *S* for each sequence, however, four *Ss* were used for each row, making 28 in all. This resulted in four replications of the same square.

The analysis of this experiment is slightly more complicated than that of a simple latin square. The total scores for the four *Ss* in each row may first be considered to form a single latin square with entries, T_{ij}, where the subscripts refer to rows and columns. Let $T_A, T_B, \cdots, T_G$ refer to total scores for the seven angles (letters); $T_{\text{I}}, T_{\text{II}}, \cdots, T_{\text{III}}$ refer to totals for ordinal position within sequences (columns); $T_L, T_M, \cdots, T_R$ refer to totals for the seven sequences (rows); and $T_1, T_2, \cdots, T_{28}$ refer to total scores for the 28 *Ss*. The sum of all squared scores is ΣX^2, and the correction factor, $C = 1/196\ (\Sigma X)^2$. The appropriate analysis is given in Table II.

The first three rows of Table II yield sums of squares which are familiar. The fourth row gives a term which would be the error estimate for a single 7×7 square. Since another error estimate is avilable, however, the usual error mean square—here called *Square Uniqueness*—can be evaluated to see whether or not it has been inflated or deflated by the unique pattern of confounding which occurred in the interactions of this particular square. If significant inflation or deflation occurs the experimenter is thus made aware of it, although he is still unable to discover which interaction or interactions are causing the difficulty. Presumably if the $SS_{\text{square uniqueness}}$ is inflated or deflated the reverse in-

TABLE II

OUTLINED ANALYSIS OF REPLICATED 7×7 LATIN SQUARE USED IN THE CORRIGAN-BROGDEN EXPERIMENT ON PRECISION OF LINEAR PURSUIT MOVEMENTS

Source of Variation	*df*	*Sum of Squares*
(1) Angles (Letters)	6	$\frac{1}{28}(T^2_A+T^2_B+\cdots+T^2_G)-C$
(2) Sequences of Angles (Rows)	6	$\frac{1}{28}(T^2_L+T^2_M+\cdots+T^2_R)-C$
(3) Ordinal Position of Angles in Sequence (Columns)	6	$\frac{1}{28}(T^2_I+T^2_{II}+\cdots+T^2_{VII})-C$
(4) Square Uniqueness	30	$\frac{1}{4}(T^2_{11}+T^2_{12}+\cdots+T^2_{17}+T^2_{21}+\cdots+T^2_{77})-C-(SS_{angles}+SS_{seq}+SS_{ord\ pos})$
(5) Individual *Ss* within Sequences	21	$\frac{1}{7}(T^2_1+T^2_2+\cdots+T^2_{28})-C-SS_{sequences}$
(6) Error	126	$SS_{Tot}-(SS_{angles}+SS_{seq}+SS_{ord\ pos}+SS_{square\ uniq}+SS_{ind\ within\ seq})$
(7) Total	195	ΣX^2-C

fluence will be present in one or more of the main effects. If the *Square Uniqueness* mean square is not significantly greater than the *Error* mean square, the two sums of squares may be combined to obtain a more reliable error estimate. The significance of the fifth mean square, that for *Individuals within Sequence*, provides a test of the reliability of the scores used (6, p. 249 f.). The *Sequences* mean square should be tested against the *Individuals within Sequences* mean square as well as against the *Error* term in order to learn whether a significant F for *Sequences* mean square divided by *Error* mean square could be accounted for in terms of significant variation between *Ss*.[7] The error term itself is free of all variation directly due to *Ss*, angles, ordinal position, sequences, and square uniqueness. It may contain some interaction variance, but this would be restricted to interactions between *Individual Ss within Sequences* and *Angles* and *Ordinal Position within Sequences*. Thus by replicating the same square a number of times, a more nearly pure error estimate is obtained.

Greco-latin Squares

An $N \times N$ greco-latin square consists of N latin letters and N greek letters in each of N rows, forming an $N \times N$ square in which each greek and latin letter occurs just once in each row and each column, and each greek letter occurs just once with each latin letter. The construction of such squares is described by Fisher and Yates (5). An example of a 5×5 greco-latin square is given below:

$D\alpha$	$A\gamma$	$C\epsilon$	$B\delta$	$E\beta$
$A\epsilon$	$C\beta$	$E\delta$	$D\gamma$	$B\alpha$
$E\gamma$	$B\epsilon$	$D\beta$	$C\alpha$	$A\delta$
$C\delta$	$E\alpha$	$B\gamma$	$A\beta$	$D\epsilon$
$B\beta$	$D\delta$	$A\alpha$	$E\epsilon$	$C\gamma$

In the greco-latin square, rows, columns, latin letters, and greek letters are orthogonal to each other. The analysis of the 5×5 greco-latin square is outlined below:

Source of Variation	*df*
Rows	4
Columns	4
Latin letters	4
Greek letters	4
Error	8
Total	24

[7] This test was found to be important in the Corrigan-Brogden experiment.

The greco-latin square is admirably suited to experiments such as that of Buxton and Ross[8] (1) in which four lists of nonsense syllables were memorized under four experimental conditions by each of 32 *Ss*. In the typical experiment of this type the lists will differ in difficulty, the *Ss* will show a practice effect, and there will be consistent differences between *Ss* in memorizing ability. Effects of all of these irrelevant variables should be eliminated from the error estimate in an efficient experimental design if the influence of the four experimental conditions is to be properly evaluated.

Simultaneous elimination of all of these extraneous sources of variation may be accomplished by using eight different 4×4 greco-latin squares, which could be arranged in a column to form a 32 row by 4 column table. Then *rows* would yield individual differences between *Ss; columns* would be ordinal position in the sequence of procedures, yielding practice effects; *latin letters* would be assigned to experimental conditions; and *greek letters* would be assigned to lists. The analysis of variance is outlined below:

Source of Variation	*df*
(1) Experimental condition (latin letters)	3
(2) Variation in list difficulty (greek letters)	3
(3) Order or practice effect (columns)	3
(4) Individual differences between *Ss* (rows)	31
(5) Error	87
(6) Total	127

The sums of squares are obtained as before. The *Individual Differences* sum of squares will be equal to one-fourth of the sum of the 32 squared individual totals minus the correction factor. The SS_{Error} will be the SS_{total} minus all other sums of squares.

This design is unusually efficient in that individual differences are effectively removed from the error estimate. Because individual differences are orthogonal to the variation produced by the experimental factor, each *S* serves as his own control through all four procedures.

One last experiment will be outlined to illustrate how the principles of the factorial experiment may be combined with principles of the greco-latin square. In this well-planned investigation (13) three experimenters ran nine monkeys through a nine-week experimental program under three conditions of motivation. Each week each monkey was trained on a different list of essentially similar problems. By use of a

[8] Buxton and Ross obtained an effective double-elimination of individual variability by means of analysis of covariance. The greco-latin square procedure described here is an alternative design which they did not actually use.

9×9 greco-latin square it was possible to eliminate the effects of experimenters, monkeys, motivating conditions, problem lists, and weeks.[9] The experimental design is presented diagramatically in Table III. In this design the successive weeks of experimentation are assigned to *rows*, the motivational conditions and experimenters are assigned to *columns*, the monkeys are assigned according to the *latin letters*, and the lists of problems are assigned according to the *greek letters*.

TABLE III

USE OF 9×9 GRECO-LATIN SQUARE IN EXPERIMENT ON ROLE OF REVERSAL LEARNING MOTIVATION IN DISCRIMINATION

Motivation		I			II			III			Row
Experimenter		Don	June	Oscar	Don	June	Oscar	Don	June	Oscar	Totals
	1	$C\zeta$	$A\delta$	$H\iota$	$D\gamma$	$B\epsilon$	$F\beta$	$G\theta$	$E\alpha$	$I\eta$	T_1
	2	$B\eta$	$C\theta$	$G\alpha$	$F\delta$	$A\iota$	$E\zeta$	$I\gamma$	$D\epsilon$	$H\beta$	T_2
	3	$G\iota$	$H\eta$	$F\gamma$	$B\zeta$	$I\theta$	$A\epsilon$	$E\beta$	$C\delta$	$D\alpha$	T_3
	4	$I\alpha$	$G\beta$	$E\delta$	$A\eta$	$H\gamma$	$C\iota$	$D\zeta$	$B\theta$	$F\epsilon$	T_4
Week	5	$H\epsilon$	$I\zeta$	$D\theta$	$C\beta$	$G\delta$	$B\alpha$	$F\eta$	$A\gamma$	$E\iota$	T_5
	6	$D\delta$	$E\epsilon$	$C\eta$	$H\alpha$	$F\zeta$	$G\gamma$	$B\iota$	$I\beta$	$A\theta$	T_6
	7	$E\gamma$	$F\alpha$	$A\zeta$	$I\iota$	$D\beta$	$H\theta$	$C\epsilon$	$G\eta$	$B\delta$	T_7
	8	$F\theta$	$D\iota$	$B\beta$	$G\epsilon$	$E\eta$	$I\delta$	$A\alpha$	$H\zeta$	$C\gamma$	T_8
	9	$A\beta$	$B\gamma$	$I\epsilon$	$E\theta$	$C\alpha$	$D\eta$	$H\delta$	$F\iota$	$G\zeta$	T_9
Col. Totals		T_a	T_b	T_c	T_d	T_e	T_f	T_g	T_h	T_i	ΣX

Monkeys designated by latin letters.
Problem lists designated by greek letters.

The columns are used in an interesting manner. Each of the three experimenters tests under each of the three motivational conditions so that the two factors, motivation and experimenter, are completely orthogonal to each other, forming a three by three factorial experiment within the larger structure of the experiment. This means that the *motivation×experimenter* interaction can be evaluated. The greco-latin square enables the experimenter to control the variation due to the extraneous factors, of: (a) individual differences between monkeys; (b) variability in difficulty of the nine lists of problems; and (c) general

[9] It should be noted that if Meyer had tried to use a 10×10 square, he could not have formed a greco-latin square. The same is true of the 6×6, 14×14, 18×18 squares, etc. For this reason, when several factors are to be controlled simultaneously in a greco-latin or hyper-greco-latin design, these squares should be avoided.

TABLE 4

OUTLINED ANALYSIS OF VARIANCE OF A 9×9 GRECO-LATIN SQUARE WITH TWO ORTHOGONAL FACTORS IN THE COLUMNS

Source of Variation	*df*	*Sums of Squares*
(1) Motivation	2	$SS_{motiv} = \frac{1}{27}(T^2_I + T^2_{II} + T^2_{III}) - C$
(2) Experimenter (Columns)	2	$SS_{exper} = \frac{1}{27}(T^2_{Don} + T^2_{June} + T^2_{Oscar}) - C$
(3) Motivation × Experimenter	4	$SS_{motiv \times exper} = SS_{columns}{}^{*} - (SS_{motiv} + SS_{exper})$
(4) Weeks (Rows)	8	$SS_{weeks} = \frac{1}{9}(T^2_1 + T^2_2 + \cdots + T^2_9) - C$
(5) Monkeys (Latin letters)	8	$SS_{monkeys} = \frac{1}{9}(T^2_A + T^2_B + \cdots + T^2_I) - C$
(6) Problem Lists (Greek letters)	8	$SS_{lists} = \frac{1}{9}(T^2_\alpha + T^2_\beta + \cdots + T^2_\iota) - C$
(7) Error	48	$SS_{error} = SS_{total} - (SS_{rows} + SS_{cols} + SS_{monkeys} + SS_{lists})$
(8) Total	80	$SS_{total} = \Sigma X^2 - C$

* $SS_{columns} = \frac{1}{9}(T^2_a + T^2_b + \cdots T^2_i) - C$

practice effects which might cause week to week improvement in the performance of the animals.

The analysis of variance with the procedures for calculating the sums of squares of the data of the experiment is outlined in Table IV. The eight df for columns is split up, two for experimenters, two for motivational conditions, and four for the interaction of these two factors. The design is highly efficient and should give a fairly pure error estimate. The purity of the error term illustrates a good feature of the larger randomly selected latin squares. The larger latin squares are rarely used in agricultural field designs because as they become large, the increased efficiency does not compensate for their high cost. For psychological work, however, the larger latin squares are just as efficient as the small ones, and, in addition, will usually yield purer variance estimates. The use of factorial design in the columns, as in the example above, or the rows, or in both, is also more convenient in the larger squares.

Concluding Remarks

This brief paper was not intended to exhaust the variety of fruitful applications of the latin square to psychological research. It is clear that whenever an experimenter is faced with the problem of balancing out effects of the temporal order in which two or more procedures are to be followed with the same *Ss* he should consider the use of the latin square principle. For further information the psychologist will find good general descriptions of the latin square in these references (**3, 4, 7, 14, 16, 17**), but specific psychological applications have as yet appeared only rarely (**6, 15**). It seems reasonable, however, to predict that the latin square design will prove just as valuable to the psychologist as the factorial experiment, and that further psychological applications of the latin square will soon be forthcoming.

BIBLIOGRAPHY

1. Buxton, C. E., & Ross, H. V. Relationship between reminiscence and type of learning technique in serial anticipation learning. *J. exp. Psychol.*, 1949. (In press.)
2. Corrigan, R. E., & Brogden, W. J. The effect of angle upon precision of linear pursuit movements. *Amer. J. Psychol.*, 1949. (In press).
3. Fisher, R. A. *Statistical methods for research workers.* London: Oliver and Boyd, 1941.
4. Fisher, R. A. *The design of experiments.* New York: Hafner, 1947.
5. Fisher, R. A., & Yates, F. *Statistical tables for biological, agricultural, and medical research.* London: Oliver and Boyd, 1938.
6. Garrett, H. E., & Zubin, J. The analysis of variance in psychological research. *Psychol. Bull.*, 1943, **40**, 233–267.
7. Goulden, C. H. *Methods of statistical analysis.* New York: Wiley, 1939.

8. GRANT, D. A., SCHNEIDER, D. E., & GOODALE, J. C. Group pre-training of serial rote learning by means of a moving picture technique. *J. gen. Psychol.*, 1949. (In press.)
9. KANNENBERG, K. M. A comparison of results obtained from the Thematic Apperception Test under two conditions of administration. M. A. Thesis, University of Wisconsin Library, 1948.
10. KOGAN, L. S. Analysis of variance—repeated measurements. *Psychol. Bull.*, 1948, **45**, 131–143.
11. LINDQUIST, E. F. *Statistical analysis in educational research.* Boston: Houghton-Mifflin, 1940.
12. MCNEMAR, Q. Sampling in psychological research. *Psychol. Bull.*, 1940, **37**, 331–365.
13. MEYER, D. R. The effect of food deprivation on discrimination reversal learning. (In preparation.)
14. SNEDECOR, G. W. *Statistical methods.* Ames, Iowa: Collegiate Press, 1946.
15. THOMSON, G. H. The use of the latin square in designing educational experiments. *Brit. J. Educ. Psychol.*, 1941, **11**, 135–137.
16. WISHART, J. Field trials: Their layout and statistical analysis. *Imp. Bur. Plant Breeding and Genet.*, School of Agriculture, Cambridge, 1940.
17. YATES, F. Incomplete latin squares. *J. agri. Sci.*, 1936, **26**, 301–315.

KINSEY'S "SEXUAL BEHAVIOR IN THE HUMAN MALE": SOME COMMENTS AND CRITICISMS[1]

LEWIS M. TERMAN
Stanford University

Like others, the reviewer has been deeply impressed by the magnitude and potential significance of Kinsey's research. From the first volume of his projected series it is obvious that no one has ever obtained so much information from so many persons regarding the most secret phases of their sexual histories. The advance publicity given the book had prepared for it a hearty welcome, and a cursory examination of its contents tended to confirm the favorable opinions previous reviewers had expressed. However, a careful reading and rereading of the report has raised so many questions that it has seemed desirable to publish the following comments in the hope that they may (1) lead others to examine the book more critically, and (2) have a beneficial effect upon the treatment and exposition of data in the volumes which are to follow.

It would be premature to attempt a general appraisal of Kinsey's entire investigation on the basis of this progress report. The comments here offered are not an all-round appraisal of even the first volume, since they have dealt almost entirely with its shortcomings and inadequacies. The reviewer has felt justified in confining his comments chiefly to the demerits of the report because its merits have been recounted so extensively by others.[2]

Scope and validity of the basic data. All of the data were obtained through personal interviews, as Kinsey has no confidence in the anonymous questionnaire. It is probably true that some of the 300 to 500 items of information called for in his interviews could not have been obtained by any kind of questionnaire with comparable accuracy. On the other hand, it is conceivable that some of the information would have been more accurately reported had a method been used which prevented the investigator from learning the identity of any of his respondents. Be that as it may, Kinsey has chosen the slower and the harder

[1] KINSEY, A. C., POMEROY, W. B., and MARTIN, C. E. *Sexual behavior in the human male*, Saunders, 1948. Pp. 804.

(As Kinsey is responsible for the general plan of the investigation, gathered most of the data, and presumably wrote the report, the reviewer has omitted the names of Pomeroy and Martin in references to authorship.)

[2] See especially: *Sex habits of American men, A symposium on the Kinsey Report*, edited by Albert Deutsch, Prentice Hall, 1948; *American sexual behavior and the Kinsey Report*, by Morris Ernst and David Loth, Greystone Press, 1948; and *About the Kinsey Report*, edited by D. Geddes and C. Curie, New American Library, 1948.

way. One can only marvel at the zeal and perseverance of an investigator who would undertake to carry through 100,000 interviews, each requiring on the average 90 minutes or more.

The amount of information obtained is surprisingly great for a single interview, covering, as it does, not only details about current sexual activities, but in equal detail the earlier activities of the subject as far back as memory can recall them. Unfortunately, the author tells us almost nothing about the wording of the questions asked, a matter which the professional pollsters have found to be extremely important. The reason given for this omission is lack of space, but since the wording of questions vitally affects the interpretation of almost every statistic in this 800-page book, the omission is regrettable.

What the author does say about the questions is not always reassuring. In the first place, we are told that they have never been standardized; instead, the manner of wording them varies according to the age, intelligence, and personality of the subject being interviewed. The necessity of alternative forms of wording will be granted, but without knowledge of the forms deemed permissible, no other investigator can repeat the Kinsey experiment with any assurance that he is getting comparable results. The two assistants (Pomeroy and Martin) had undergone a year of training in interview methods, yet each obtained results which at various points differ reliably from those of the other and from Kinsey's. This is true despite the fact that sex, race, marital status, age, education, religion, and rural-urban residence were held constant for the three groups of interviewees compared.

Consider, for example, the mean frequencies found by the three interviewers in the age groups adolescence to 15, and 16–20. The means given (p. 134) are for total outlet, masturbation, nocturnal emission, premarital coitus, and homosexual contact. This gives ten comparisons of means in the two age groups for the total population and ten additional comparisons for the *active* population (that is, the population which has practiced a given sexual activity). In nine of the ten comparisons for the *total* population the Kinsey mean is highest, and in most cases reliably so. For premarital coitus in the age group 11–15, Kinsey's figure is three times that of Pomeroy and more than twice that of Martin. For homosexual contact at this age Martin's figure is less than one-fourth as high as either Kinsey's or Pomeroy's. In age-group 16–20 Kinsey's figure for premarital coitus is nearly twice that of Pomeroy, and for homosexual contact it is four times that of Martin. Differences of similar magnitude are found in the ten sets of mean frequencies for the *active* population.[3]

[3] The generally higher frequencies found by Kinsey are said to be due to the fact that some of the more promiscuous and difficult subjects were assigned to him for interview. However, we learn from Table 22 and Table 23 that there are reliable differences between Kinsey's own data of 1938–1942 and his data of 1943–1946.

A questionable feature of the interview technique is the author's practice of always placing the burden of denial upon the subject. The practice is defensible with a majority of subjects, but it could easily invalidate the reports of children and of feeble-minded or low-level adults. "We always assume," says Kinsey, "that everyone has engaged in every type of activity. Consequently we always begin by asking *when* they first engaged in such activity" (p. 53). Yet, in the second paragraph preceding this statement the author warns that "In his tone of voice and in his choice of words the interviewer must avoid giving the subject any clue as to the answers he expects." On p. 55 the author describes a technique for what he calls "proving the answer," which is said to be useful with uneducated persons and the feeble-minded. The method is "to pretend that one has misunderstood the negative replies and ask additional questions, just as though the original answers were affirmatives . . . . " Example. "Yes, I know you have never done that, but how old were you the *first* time that you did it?" Anyone familiar with the experimental literature on suggestibility will wonder about the possible effects of this technique.

The author is little concerned about the danger of fabrication; that, he believes, can be taken care of by "Looking an individual squarely in the eye, and firing questions at him with maximum speed . . . " (p. 54). Cover-up, the author says, is harder to catch, and its possible influence on the incidence figures obtained is specifically mentioned in a number of places. For example, on p. 499 he says that "while college men more often admit their experience [of masturbation], there are males in some other groups who would admit almost any other kind of sexual activity before they would give a record of masturbatory experience." Elsewhere he states that histories on socially taboo items generally are difficult to secure from older married subjects of superior levels. On p. 54 it is said that cover-up is combatted by "the use of a considerable list of interlocking questions which provide cross-checks throughout the history, and particularly in regard to socially taboo items." However, the author is not very explicit about the exact nature of these cross-checks, and the examples given do not impress this reviewer as altogether convincing.

An additional source of error is the long-distance memory report. The author tries to check on the extent of such errors by retakes on 162 subjects and by noting the extent of husband-wife agreement in 231 married couples who gave memory reports on the same facts. The retakes were made after intervals ranging from 18 months to 7 years (mean, 38.5 months). Table 13 shows the take-retake correlations to be very high for vital statistics data and for percentages who had engaged in specified sexual activities. On frequencies of outlet, the correlations were much lower, ranging from .58 to .67, and in most cases they were similarly low on reported age when given sexual activities first oc-

curred. For example, on age at first ejaculation the two reports disagreed by more than one year for 52.4 per cent of subjects ($r = .58$). On reported age at first nocturnal emission 76.2 per cent of subjects showed disagreement of more than a year ($r = .54$). In general, the correlations on frequencies and on age at first experience are not high enough to permit very reliable comparisons with other variables. Moreover, as the author admits (p. 125), the retakes do not test the validity of the data, but rather the constancy of memory and of tendency to cover-up.

Regarding the comparisons of reports by the 231 married couples, Kinsey says "the record shows an amazing agreement" between the statements of husbands and wives. However, examination of Table 14 shows that most of this "amazing agreement" is on 12 items of vital statistics, such as number of years married, length of premarital acquaintance, length of engagement, age at marriage, number of children, amount of education, occupation of father, etc. The husband-wife correlation was only .50 for average frequency of coitus in early marriage, .54 for maximum frequency, and .60 for average frequency "now." Of the 14 correlations which are .80 or higher (out of the total of 32), 10 concerned vital statistics and the others concerned the practice of coital foreplay and intercourse in the nude.

Unable to correct for errors of memory, the author proceeds statistically as though they did not exist; that is, he gives the same weight to reports based upon remote recall as he gives to reports of current activities. In the computation of mean frequency of masturbation at age 15, for example, the memory report of a 50-year-old counts as heavily as the report of a 15-year-old.

The validity of Kinsey's sampling. The problems of sampling in an investigation of this kind are of paramount importance. If every response by every subject were completely truthful, the resulting data could still be misleading if the groups interviewed were not representative of their kind. Indeed, representativeness is incomparably more important than sheer numbers, for however numerous the subjects interviewed, if the sampling is biased the generalizations will be biased. An advertising circular from the publisher states that the interviews were "conducted with full regard for the latest refinements in public opinion polling methods." However, Kinsey's discussion of his sampling procedure on p. 92 ff. makes it clear that no scheme of randomization like those common in public opinion polls has been used in this study. He depended instead on huudred-per-cent samples of certain groups and upon diversification of the total population by the addition of whatever subjects were available for interviewing. The latter were subjects who volunteered or could be persuaded to cooperate, and are said (p. 95) to constitute 74 per cent of the 12,000 males and females who have been interviewed to date.

Kinsey is not to be criticized for not using the methods common in

public opinion polls; as he points out, a strictly random selection of subjects in a study of sexual behavior would not have been feasible. The report is open to criticism, however, for not giving us the information needed to judge the representativeness of either the volunteers or the hundred-per-cent samples. The *N*'s of contributing groups are almost never stated. Hundred-per-cent samples were obtained from 62 groups, of which 42 were of college level. Seven of the remaining 20 groups were institutional cases in four delinquent groups, two penal groups, and one group in a "mental" institution. There were three classes of junior high school students, three speech-clinic groups, three rooming-house groups, two groups of conscientious objectors, a group of N.Y.A. workers, and a group of hitch-hikers. Whatever the *N*'s of these individual groups may have been, it is unlikely that the total hundred-per-cent sample could have been representative of the U. S. population, or could have been made so by any kind of statistical doctoring.

The information given about the volunteers is equally incomplete. We are told (p. 38) that about half of the 12,000 histories to date have been obtained through contacts resulting from several hundred lectures by the author to groups numbering in all perhaps 50,000 persons. We do not know how those who attended the lectures differed from those who might have attended but did not, nor how the 6,000 who heard the lectures and allowed themselves to be interviewed differed from the 44,000 who heard them but did not cooperate. The author lists (p. 39) 32 groups of "contact" persons, numbering "many hundred" in all, who helped in obtaining volunteers. Seven of these 32 were delinquent groups: male prostitutes, female prostitutes, bootleggers, gamblers, pimps, prison inmates, thieves and hold-up men. These, presumably, would have brought in others of their kind, but in what numbers they did so we are not told. Elsewhere (p. 15) the author lists a dozen prison populations which, he says, "have augmented our understanding of economically and educationally lower social levels, and of the broken marriages which are in the histories of a high proportion of the penal inmates." Additional institutions mentioned which were not penal or correctional included a state school for feeble-minded, two children's homes, and two homes for unmarried mothers. On p. 16 we learn that subjects were obtained from "homosexual communities" in Chicago, New York, Philadelphia, Indianapolis, and St. Louis; also from "underworld communities" in Chicago, Peoria, Indianapolis, New York City, and Gary (Indiana).

On p. 392 Kinsey states that he has data on more than 1,200 persons who have been convicted of sex offenses. We are not told how many of the convicted sex offenders are included in the population of 5,300 white males for whom data are summarized in this volume. On p. 210 we learn that data on frequencies of penal groups while in prison were not included in the frequency calculations, but their memory reports of

sexual activities prior to their imprisonment were presumably included along with the data from other subjects. The scanty and scattered information available warrants the suspicion that Kinsey's educationally low-level groups may have been far from typical of this level in the generality. The suspicion is strengthened by the information (p. 213) that 49.4 per cent of the underworld males have a mean frequency of outlet that is equalled by only 7.6 per cent "of any population."

Similar questions arise about the representativeness of nearly all the sub-groups, except possibly the group which had attended college one or more years, and which, incidentally, made up more than half of his 5,300 white males. Little information is given as to the source of the 9–12 educational group, and even less about the source of the rural group. The rural group is vaguely defined (p. 451) as including those subjects who spent any "appreciable portion of the years between 12 and 18" on an operating farm. Specific mention is made of interviews carried out over a period of years in certain "remote" and "isolated" rural communities, but what they were like, or how many subjects they furnished, we are not told. It is unfortunate that the author did not adhere to the excellent rule which he laid down on p. 33 to the effect that "Each segment which is studied must be precisely delimited, and all conclusions must be confined to such precisely defined groups."

Other fragments of information that have a bearing on the sampling are found throughout the book. On p. 544 mention is made "of the six thousand marital histories in the present study, and of nearly three thousand divorce histories . . . . " These, presumably, are in the twelve thousand histories collected to date from males and females. Surely a population in which the divorce histories are half as numerous as the marital histories provides a shaky foundation for a census of sexual behaviors.

One question regarding the representativeness of Kinsey's sampling is whether the subjects who volunteered, and who account for about three-fourths of his total population, tended to be of a special sort. One might suppose that persons most willing to talk about their sex lives would be, in a disproportionate number of cases, those least inhibited in their sexual activities. On p. 37 Kinsey says that many who volunteered did so because they were seeking information or help in connection with their personal problems. The best way to check on the representativeness of the volunteer sample would have been to compare it directly with the hundred-per-cent sample. Kinsey does not do this, but he does (on pp. 94–102) compare the hundred-per-cent sample with what he strangely calls the "partial sample," the partial sample being defined in a footnote as the hundred-per-centers plus the volunteers. That is, his comparisons are really between the hundred-per-cent sample and his *complete* sample.

The results of these comparisons are summarized in Table 3, which

gives for single males of college level the mean and median frequencies of six types of sexual outlet for three separate age groups: adolescence to 15, 16–20, and 21–25. The figures show that at all age levels, and for all kinds of outlet except nocturnal emissions, the means run consistently higher for the complete sample. Many of the differences are very large and highly reliable despite the fact that the complete sample includes also his hundred-per-cent sample. From the data given in the table Dr. Quinn McNemar has computed the means separately for the subjects who volunteered. When these are compared with the means of the hundred-per-cent sample the differences are of course greater than those shown in the table. For premarital coitus the difference in mean frequency by this method of comparison is nearly 2 to 1 at adolescence to 15 (actual means .09 and .05), somewhat less than 2 to 1 at ages 16–20 (actual means .31 and .18), and about 3 to 2 at ages 21–25 (actual means .50 and .36). For homosexual contacts the difference as thus computed becomes nearly 2 to 1 at adolescence to 15, more than 3 to 1 at 16–20, and exactly 4 to 1 at 21–25. Differences of such magnitude confirm the suspicion that willingness to volunteer is associated with greater than average sexual activity. And since the volunteers account for about three-fourths of the 5,300 males reported upon in this volume, it follows that Kinsey's figures, in all probability, give an exaggerated notion of the amount of sexual activity in the general population.[4]

Notwithstanding the fact that all but a small fraction of the subjects interviewed resided in five mid-western states, five middle-Atlantic states, and two New England states, Kinsey "corrects" his findings (for such factors as education, age distribution, marital status, rural-urban residence, et cetera) to show incidences and frequencies for various subgroups on each kind of sexual outlet in the entire U. S. population. It is evidently the intention of the author to spread his interviews in the years to come more or less equally throughout the country in proportion to the population of the individual states, but even when this is done there will still remain the problem of obtaining in each area a representative sampling for each of his twelve major break-down groups. In view of the inadequacies of the sampling to date, the "corrections" to show hypothetical incidences and frequencies for the total U. S. population seem to this reviewer indefensible at the present stage of the investigation.

One of the most puzzling omissions in the book is the author's failure to give the complete age distribution of his subjects at the time they were interviewed. Mention is made of subjects who were in their 70's and 80's, but because of small *N*'s in the upper age brackets the data for

[4] This conclusion is based on the comparisons between the volunteers and the hundred-per-centers of college level. The author does not compare the two kinds of samples for the 0–8 or 9–12 educational levels because of the relatively small N's in these groups.

most types of outlet are not summarized for age groups above 40 or 45 years. The reviewer has found no statement about the lowest age limit of the younger subjects; incidences and frequencies are summarized for some types of outlet down to the age of 8 years, but we are not specifically informed whether any data obtained from 8-year-olds (or for that matter from 9-, 10-, 11-, or 12-year-olds) have been reported for these 5,300 males. As previously stated, the author throws together the memory reports and the reports of current activities, giving the two kinds of data equal weight. It would have been helpful if he had shown in the tables what proportion of the *N* at any given age level was accounted for by subjects at or near that age. This proportion would be high in the late teens and early twenties, because of the large (though unstated) number of students who were interviewed while they were attending college; but for all we know, the proportion may be zero for ages 8 to 12 or even later, that is, the data for the lower ages in the tables may all be based on the memory reports of older subjects, many of them 20, 30, or 40 years older.

The sample-size experiment. Kinsey gives the results of a series of drawings of random samples from his total population which were intended to establish the size of sample needed to give stable incidences and frequencies for the different kinds of sexual activity. The reviewer asked Dr. Quinn McNemar to check over the statistical procedures used in this experiment. His report follows.

In all, 40 pages (82–92; 736–765) are devoted to determining, empirically, what size of sample for each sub-group is desirable and adequate. This empirical approach involved drawing sub-samples of varying sizes from available total samples (groups) and noting what *N*, in general, led to values for means, medians, modes, modal frequencies, and ranges which were within five per cent of the *magnitudes* of the respective values for the total groups. For incidence percentages, the criterion was that sub-sample percentages should fall within two per cent of the *magnitudes* of the respective total group percentages. This procedure, which required a total of 4,279 comparisons of sub-samples with total sample values, involves some serious fallacies.

(1) Failure to recognize the fact that the sampling stabilities of means, medians, and modes are not a function of their magnitudes, but rather of trait variability. Means, for example, of the same size can, and do, have differing standard errors for constant *N* when "score" variation differs.

(2) Failure to consider the fact that these three statistics differ markedly from each other in their sampling errors even when computed from the same distribution or from distributions of similar variability. For a normally distributed trait 57 per cent more cases are required to obtain a median with a standard error equal to that of the mean. In general, the shape of a distribution affects the relative sampling stability of the median and mean.

(3) Failure to observe the fact that the sampling stability of percentages is not a linear function of their magnitudes but rather of their degree of remoteness from 50 per cent. For constant *N*, the standard error of 90 per cent is the same as the standard error of 10 per cent, whereas the criterion of 2 per cent of mag-

nitude would, if correct, imply that the error for 90 per cent is 22 times (5% of 90% divided by 2% of 10%) as large as the error for 10 per cent!

(4) Failure to note that convergence of sub-sample values to total group values must be more rapid when sub-samples are drawn from small (finite) groups than when drawn from larger groups. For instance, one would expect a sub-sample of 400 drawn from a total group of 481 to yield a value very near the total value, and likewise for 300 from 375, but a sample of 300 or 400 drawn from 1,513 or from 2,762 leaves room for considerable variation. The ignoring of this fact led to the puzzling conclusion (p. 85) that sample sizes greater than 400 "fail to show any consistent improvement," as one would expect "by standard statistical theory." Now it happens that two-thirds or 440 of the 668 samples of size 400 are sub-samples which include from 66 per cent to 87 per cent of the total cases from which they were drawn, whereas none of the samples of size 600 exceeds 60 per cent of the supply and two-thirds of the samples of 600 include less than 40 per cent of the supply. Under these circumstances and in light of that part of standard statistical theory concerned with sampling from finite universes, one would expect samples of 400 to appear more adequate than samples of 600.

In brief, incognizance of four elementary statistical principles renders worthless this elaborate effort to determine how large N should be for a sub-group.

Effects of early and late puberty. Some interesting material is summarized on the relation of sexual activity to age at onset of adolescence. The criteria used to establish this age are described as follows on p. 299.

> In the present study, the time of onset of adolescence has been fixed as the date of the first ejaculation, *unless there has been evidence that ejaculation would have been possible at an earlier age if the individual had been stimulated to the point of orgasm.* When the year of first ejaculation coincides with the year in which the first pubic hair appears, and with the time of onset of rapid growth in height, and/or with certain other developments, there is no question that that year may be accepted as the first year of adolescence. Eighty-five per cent of all male histories fall into this category. On the other hand, if the first ejaculation follows these other events by a year or more, and *if it is clear that there was no test of the individual's sexual capacity prior to the first ejaculation, and if there seems to be no question of the reliability of the memory in regard to the dates of the other adolescent developments,* then the age of onset of adolescence is better established by events other than ejaculation. When first ejaculation occurs as a nocturnal emission, it usually (though not always) does not come until a year or more after the appearance of the other adolescent developments, and the onset of adolescence should be set a year or more before the first ejaculation. (Italics by reviewer.)

Apart from the fact that the rules here set forth are involved and "iffy," the reviewer doubts whether any man ten or twenty years beyond adolescence could give more than a wild guess as to his age at first ejaculation, at first appearance of pubic hair, or at onset of rapid growth. It is not merely a matter of memory; each of these signs of adolescence makes its appearance gradually. Consider the "unless"

clause in the first sentence. What would be good evidence that ejaculation could have occurred earlier if suitable stimulation had been present? And how could the interviewer be sure whether there was or was not a test of the subject's sexual capacity prior to first ejaculation, or whether the subject's memory in regard to the other signs of adolescence is or is not reliable? Such judgments would seem to call for a kind of occult insight that most people don't have.

It is surprising that memory reports of such events, erroneous as they often must be, should reveal significant differences between early- and late-maturing subjects for both frequencies and incidences. On p. 307 it is stated that "The effect persists throughout the lives of the married males, as far as data are available." However, the *N*'s are so low at the later years that for the educational levels 0–8 and 9–12 no data on this variable are given beyond age 25. By combining the three educational levels the author is able to carry the data, with some fairly satisfactory *N*'s, to age 40. It should be noted, however, that the mixing of educational levels is a procedure which he severely criticizes in others. The figures on p. 306 show that in frequency of total outlet the ratio of early- to late-maturing married males is more than 2 to 1 at age 16–20 and about 3 to 2 at 21–25 and 26–30, but that thereafter the ratio fluctuates in the neighborhood of 1 to 1. Accordingly, the author's generalization about the persistence of the difference throughout the married life of the subjects is hardly warranted by the data presented.

The author suggests that both early puberty and continued high frequency of outlet are probably functions of the general metabolic level, although no metabolism tests of his subjects are reported. In line with this is his belief that the "early-adolescent males are more often the more alert, energetic, vivacious, spontaneous, physically active, socially extravert, and/or aggressive individuals in the population," and that, conversely, the late-maturing are more often "slow, quiet, mild in manner, without force, reserved, timid, taciturn, introvert, and/or socially inept . . . " (pp. 325–326). He bases this conclusion on personality ratings recorded at the time the interviews were made. Psychologists who have found it so difficult to devise reliable measures of such personality traits will be interested in these ratings.

Another of Kinsey's conclusions (p. 325) is that the sexual capacities do not seem to be impaired as a result of their early initiation and continued high frequency in the early-maturing male. On p. 323 he reports for 69 older impotent males a *plus* correlation of .30 between age of onset of adolescence and the age of onset of impotence, but he interprets this coefficient as indicating "that there is in actuality no significant correlation" and that impotence is as likely to occur by a given age among the early-maturing as among the late-maturing. However, the correlation is significant at the .01 level, by the small-sample technique, which suggests a definite tendency for the earlier-maturing and high-frequency males to become impotent earlier.

The stability of sexual patterns in two generations. The author gives extensive data on the stability of sexual patterns from one generation to another. He divided the entire male sample into two groups. One group included those who were 33 years of age or older at the time they contributed their histories; its median age was 43.1 years. The other group included all who were younger than 33, and its median age was 22.2 years. The two groups have been compared on incidence and frequency of every type of outlet for each age from 8 to 33 or 34 years. The author's conclusion from these comparisons, as stated on p. 397, is that "In general, the sexual patterns of the younger generation are so nearly identical with the sexual patterns of the older generation in regard to so many types of sexual activity that there seems to be no sound basis for the widespread opinion that the younger generation has become more active in its sociosexual contacts." The critical reader will want to scrutinize the tables carefully to learn the extent to which the data justify this conclusion.

Note first that the subjects in the two groups do not all belong to separate generations. Their age distributions are adjacent, and many of each group must have differed in age only a few months to a few years from many in the other group. We can be sure, therefore, that whatever differences the figures yield would have been much greater if there had been an age gap of 10 or 20 years between the groups. Secondly, it should be noted that the reports by the older group are based on much more remote recall,—on the average some 20 years more remote.

With these limitations of the data in mind, we turn next to the tabulated reports of the subjects. At the educational level 13+ we find (p. 400) that the incidence of premarital intercourse is considerably higher for the younger group from age 19 to 24 inclusive; for most of these ages it is one-fifth to one-seventh higher. At the 0–8 educational level the accumulative incidences for most ages run reliably higher in the younger group for all types of outlet except intercourse with prostitutes. At this level the incidence of premarital intercourse is four times as high for the younger group at age 12, three times as high at age 13, one-half higher at 14 and 15, and one-third higher at 16. At this educational level the incidence of petting at ages 12 and 13 is three times that reported by the older group, and at 14 it is nearly twice as high. In this 0–8 group the incidence of masturbation runs significantly higher in the younger group from 11 to 30 years; at 12 and 13 it is close to twice as high. In both the 0–8 and the 9–12 educational levels the incidence of homosexual contacts runs from one-and-one-half times as high to nearly twice as high for the younger generation, and among married men of these two levels the incidence of extramarital intercourse is in all groups between one-third and one-half higher for the younger generation.

The outlet frequencies (p. 410) show still greater differences between the two generations. For example, premarital intercourse in the 0–8 group has a mean frequency about twice as high in the younger genera-

tion as in the older. In the 9–12 group the difference is less but is statistically significant. Homosexual frequencies at both the 0–8 and 9–12 levels run consistently close to twice as high for the younger generation. Among married males of the 0–8 educational level, extra-marital intercourse occurs in the younger generation from two to five times as frequently as in the older generation.

In view of the above differences the author's assertions that "These comparisons of the sexual activities of older and younger generations provide striking evidence of the stability of the sexual mores" (p. 414), and that "There has not even been a material increase or decrease in the incidences and frequencies of most types of activity" (p. 415), are much too sweeping.

The influence of occupational level. Kinsey presents considerable data on incidences and frequencies for subjects who have moved up or down from the parental occupational class. On p. 419 he summarizes this material as follows: "In general, it will be seen that the sexual history of the individual accords with the pattern of the social group into which he ultimately moves, rather than with the pattern of the social group to which the parent belongs. . . . " Examination of the tables reveals that there are many exceptions to this rule. The data presented graphically on p. 444 are not typical of all ages, all occupational groups, or all incidences and frequencies.

It is largely the material on vertical mobility that leads Kinsey to the generalization, frequently reiterated, that a subject's life-long patterns of sexual behavior are well established before the age of 16. An examination of the figures for the various kinds of activity at successive age levels supports this generalization only in part. There is some tendency for the subject who moves to a high occupational class from a lower parental class to show by the age of 15 the types of sexual behavior characteristic of the higher class, but the extent to which this is true varies with the different kinds of outlet; and even for a particular type of outlet the rule does not hold equally well for migrants from all parental classes.

A factor which could invalidate the author's data on this issue is the possibility that subjects in the various social classes may not report with equal accuracy their sexual activities during the early teens. That this factor may have entered is suggested by the mean frequencies of nocturnal emissions (pp. 424–425). Why, for example, should subjects of parent-class 4 (skilled laborers) have only one-sixth as frequent nocturnal emissions before age 16, if they stay in the parent class, as they have if they are destined to move up to the professional class? And why should the subjects of parent-class 3 (semiskilled labor) have them only one-third as frequently before 16, if they stay put, as they have if they will later enter a profession? Such differences in a type of sexual behavior that is non-volitional render suspect all of Kinsey's data on outlet as related to occupational mobility.

The influence of religion on sexual behavior. In several passages Kinsey seems to regard the sex drives as forces of nature that will find their outlet regardless of measures taken to curb them. On p. 269 he says: " . . . it is clear that there is a sexual drive which cannot be set aside for any large portion of the population by any sort of social convention." In this connection the data presented on the sexual activity of active and inactive Protestants, of devout and inactive Catholics, and of orthodox and inactive Jews are interesting. The value of these comparisons is somewhat limited by the small *N*'s for Catholics and Jews, and by the author's failure to describe adequately his method of classifying a subject as religiously active or inactive, but in spite of these limitations the data suggest that religious attitudes have a considerable influence on most types of sexual activity. In almost every comparison the religiously active groups have lower incidences and frequencies than the religiously inactive groups of the same denomination. Unless religion merely attracts persons of low sex drive (which is doubtful), it would seem that religious attitudes exert a definitely restraining influense.

Generalizing beyond the data. As previously noted, Kinsey not infrequently makes unqualified statements which go beyond his data. These are of two kinds: (1) broad generalizations based upon small *N*'s, or upon groups which for other reasons are doubtfully representative; and (2) generalizations which are contradicted by the data given. Six additional examples are here brought together.

1. On p. 567 Kinsey asserts, in bold type, that "Not more than 62 per cent of the upper level male's outlet is derived from marital intercourse by the age of 55." On checking back to Table 85, p. 348, we find that there were only 81 upper-level married men above the age of 45 years for whom data on source of outlet are given. From Table 56, p. 252, we find that there were only 109 married men in the total population (all educational levels combined) of ages 51–55, and only 67 above the age of 55. Surely bold type is hardly suitable for sweeping conclusions based on such limited populations.

2. "Of all religious groups they [the orthodox Jewish males] are the sexually least active, both in regard to the frequencies of their total sexual outlet, and in regard to the incidences and frequencies of masturbation, nocturnal emissions and the homosexual" (p. 485). The author attributes the relatively low sexual activity in this group to "the pervading asceticism of Hebrew philosophy (p. 486), and in other passages he blames this ancient Jewish asceticism for the unrealistic severity with which most of the Christian peoples condemn departures from the Talmudic ideals. Probably very few readers will examine Kinsey's tables closely enough to discover that this interpretation is based on an *N* of only 59 orthodox Jews in the entire U.S., all of college level!

3. "Among the males who remain unmarried until the age of 35, almost exactly 50 per cent have homosexual experience between the beginning of adolescence and that age" (p. 623). This statement has been quoted by reviewers without any question of its validity; they have not taken the trouble to find out that it is based on an *N* of 68 for the 0–8 educational level, less than 50

for the 9–12 level, and 71 for the 13+ level (Tables 141, 142, 143).

4. "The condemnation of petting on the ground that it may lead to something that is worse is quite unfounded, for there is no evidence that the frequency of premarital intercourse has increased during recent generations . . ." (p. 541). As we have previously shown, only for the males of college education is Kinsey's statement approximately true, and even here the incidence figures for premarital intercourse run appreciably higher for the younger group from age 19 through 24.

5. On p. 392 it is stated that "The persons involved in these [various illicit] activities, taken as a whole, constitute more than 95 percent of the total male population. Only a relatively small proportion of the males who are sent to penal institutions for sex offenses have been involved in behavior which *is materially different from the behavior of most of the males in the population.*" Italics by the reviewer.) Even if the first of these statements were true, there would still be reason to doubt the validity of the second. It is as though one said that if 95 percent of all males have at some time in their lives stolen something, those who are sent to penal institutions for theft or burglary are not materially different from most males in the population.

6. On p. 387, speaking of the eighth grade teacher's violent reaction on discovering that a boy in her class has had intercourse with one of the girls, Kinsey says: "The teacher does not realize that more than a fourth (28%) of all her other eighth grade boys have similarly had intercourse." This statement is misleading. Table 136 which he cites in support of his statement merely tells us that of about 680 *adult* males of all ages (and of unspecified origin) who did not go beyond the eighth grade, 28 percent admitted having had intercourse before the age of 15 years. Again one would like to know who these adult males were who had only 0–8 grades of schooling.

Judgments of evaluation or interpretation. Although Kinsey has more than once asserted that his job is to discover and to report facts, he nevertheless does not hesitate to express judgments of evaluation and interpretation for which no data, or only inadequate data, are given. Such judgments often appear to be based upon nothing more than vague impressions or intuitions. Some of them are closely akin to moral evaluations, although Kinsey time and again disavows any intention to pass moral judgment or any competency to do so.

1. On p. 211 the author passes severe judgment on the 57 psychologists and 70 psychiatrists who are said to be "prejudiced" against masturbation, homosexuality, and extra marital intercourse. "These [prejudices] are all rationalizations, clutched at in support of a sexual suppression that is too often taken for sublimation."

2. Regarding the 179 males of lowest frequency in his population, the author says (p. 209) that 52.5 percent of them were "apathetic," and that they "never, at any time in their histories, have given evidence that they were capable of anything except low rates of activity." On p. 211 we are told that 58 percent of these 179 cases were "timid or inhibited individuals—afraid of their own self condemnation if they were to engage in almost any sort of sexual activity." The author adds that "some of these individuals become paranoid in their fear of moral transgression, or its outcome," and that 9 of them had attempted

suicide. These alleged effects of sexual restraint, presented as if they were typical, could well frighten the too chaste youth into greater sexual activity.

3. We read on p. 525 that: "On the whole, the males who are most dependent upon nocturnal emissions are those who are slow in developing physically, those who are slow in their nervous reactions or unresponsive to the usual sexual stimuli, or those who are timid and awkward in making social contacts. They are the males who are most often restrained for moral reasons. There are some outstanding exceptions to this, proving that a multiplicity of factors may be involved in determining the frequencies of nocturnal emissions; but, *by and large, emissions are most often depended upon by the male who has not made what the psychiatrist would call a good socio-sexual adjustment.*" (Italics by the reviewer.) Here, obviously, Kinsey is reporting a personal impression which may or may not accord with the facts, and the language he uses suggests that a youth who wants to escape socio-sexual maladjustment would do well to supplement his nocturnal emissions by some other kind of outlet—presumably sexual intercourse, since masturbation usually lacks the social element.

4. On p. 383 Kinsey says: "As a matter of fact, a male who has been so restrained [he is talking about the college-educated male] often has difficulty in working out a sexual adjustment with his wife, and *it is doubtful whether very many of the upper level males would have any facility in finding extra marital intercourse, even if they were to set out deliberately after it.*" Italics by the reviewer.) However, the data given in Table 85, p. 348, show that it is precisely the upper level male who, as he gets older, finds a larger and larger proportion of his intercourse outside of marriage, whereas the opposite is true of the 0–8 and 9–12 education levels.

5. "For the boys who have not been too disturbed psychically, masturbation has, however, provided a regular sexual outlet which has alleviated nervous tensions; and the record is clear in many cases that these boys have on the whole lived more balanced lives than the boys who have been more restrained in their sexual activities" (p. 514). The judgment here about "balanced lives" is primarily psychological or psychiatrical, but it is doubtful whether many psychologists or psychiatrists would render so positive a judgment, pro or con, on this issue after a single interview.

6. On p. 661 the author, apropos of an individual's preference for a sexual partner of the same or opposite sex, says that "This problem, is after all, part of the broader problem of choices in general: the choice of the road that one takes, of the clothes that one wears, of the food that one eats, of the place in which one sleeps, and of the endless other things that one is constantly choosing." That it is a problem of choosing is evident enough, but to many a reader the implication of the passage will be that the sex chosen as partner in a sexual activity is as unimportant as one's preferences regarding food or the cut of one's clothes.

At this point the reader may protest that it is unfair to quote these passages out of their context. So far as the individual passages are concerned, the reviewer does not think that their meanings have been seriously distorted. In one respect, however, the list of quotations *is* unfair; namely, in the fact that passages scattered through several hundred pages of the book are here brought together and laid end to end.

Admittedly they are not random samples of the author's generalizations and evaluations—no man of science would be consistently so reckless in his use of language. Nevertheless, the passages quoted are just the ones that are most likely to impress the youth who is in search of authoritative justification for the unrestrained satisfaction of his sexual urges.

Summary

Insufficient information has been given about the specific questions asked in the interview, and about the precautions taken to reduce cover-up and to eliminate the possible effects of suggestion. For this reason it would be difficult if not impossible for anyone to repeat the experiment and obtain comparable results.

The author has failed to give adequate information about the sources from which he obtained the various segments of the total population interviewed. From the facts given, we are not able to judge the representativeness of any of the sub-groups; the rural, the urban, the occupational classes, the 0–8 or the 9–12 educational levels, or the different religious classes. It is especially regrettable that no definite information has been given about the number of subjects among the 5,300 males interviewed who were obtained from penal or mental institutions, from underworld sources, or from homosexual groups. We are not told the proportion of divorces to marriages among those in the 5,300 who were or had been married. We are not even given the complete age distribution of subjects at the time they were interviewed.

From McNemar's criticisms of the sample-size experiment it is evident that Kinsey needed more statistical guidance than he has had. It is a pity that the 40 pages wasted on this experiment were not devoted to specific information about the nature and source of the individual groups that made up the total population.

The statistical "correction" of the found data for the purpose of showing incidences and frequencies for the total U. S. population of given age, educational level, or other class is a dubious procedure at the present stage of the investigation, dubious both because of the small N's in many sub-groups and because of their doubtfully representative nature.

No distinction has been made between reports based upon remote recall and reports of current or near-current activities. The two kinds of data have been thrown together and given equal weight. Nowhere are we told what proportions of the data for a given age are based upon remote recall and upon report of recent events. This procedure increases the N's for given activities, but at what cost in reduced validity of data no one can say.

In his text Kinsey has made many sweeping factual statements which are only weakly supported by the statistical tables, or which, in certain instances, are contradicted by the tables. Other statements are made as though they were factual when apparently they are based upon nothing more than personal impressions.

Notwithstanding Kinsey's frequent reiteration that his job is to report facts rather than evaluations, it has been possible to quote numerous passages in which recklessly worded and slanted evaluations are expressed, the slanting being often in the direction of implied preference for uninhibited sexual activity.

The reviewer fully agrees with Kinsey that the facts about human sexual behavior should be brought to light, and he regards this investigation as so important that he sincerely hopes it can be carried through to completion. But he also hopes that the many faults of exposition and interpretation to be found in this volume will not be repeated in those which follow.

BOOK REVIEWS

Assessment of men—selection of personnel for the Office of Strategic Services. The OSS Assessment Staff. New York: Rinehart, 1948. Pp. xv+541.

The development of the Office of Strategic Services covered the initiation and implementation of the intricate, highly diverse, full-scale type of intelligence service necessary for a modern nation at war. As with the other military services, it developed that the ordinary recruiting procedures were not sufficiently selective and needed supplementation by some program of personality assessment if efficient personnel were to be obtained. As a result an assessment staff was instituted late in 1943. This book sets forth clearly and in great detail the selection procedures which that staff utilized. It is well written, effectively illustrated, and presented in pleasing format. The program set forth is "organismic" rather than "elementalistic," designed to reveal the dynamic organization of the total personality. As such it relies strongly on observation of the individual performing under stress in group situations, and derives directly from the methods used by Simoneit in Germany and the War Office Selection Board program for officer candidates in the British Army.

The book not only describes these procedures but attempts a justification of their use. Its description is eminently successful, its justification is less so. In reading the book one must carefully set aside the question of whether these assessment measures sound interesting, stimulating, and revelatory of personality, and ask instead whether their successful application to a special selection task has been objectively demonstrated by the data set forth. There is no final answer in this book.

The assessment program was instituted without provision for any adequate criteria against which its operating efficiency could be evaluated. Such criterion data as were gathered later were fragmentary, impure, and of little scientific value. The force of this criticism is somewhat disarmed, however, by the complete honesty and candor with which the authors recognize and stress this weakness.

For this reason, the critical reader will find so many points that he might wish to attack that a careful discussion of them all becomes impossible within the confines of a book review. One problem may serve as an illustration. In justifying their assessment techniques the authors compare the rate of neuropsychiatric breakdowns among unassessed as against assessed personnel. They state that the precise total of OSS personnel was never determined, and use as a base figure a number "which is probably correct within 10 per cent." No mention is made of the facilities for recognizing neuropsychiatric conditions or of the availability of channels for their disposal, although both these factors influence the

validity of any recorded figures for incidence. Moreover, one assumes that the assessed group is temporally a later group, and we know that length of service is a factor in producing maladjustment. Finally, we must assume that since the initiation of a screening program came after the recognition of a personnel problem, general recruiting procedures also were improved and the group available for assessment must have been more highly selected than the earlier recruits.

Two general criticisms are valid and may be raised in the form of questions:

1. Was a determined effort made and genuinely fought through administrative channels to provide for evaluative criteria at the initiation of the program?
2. In view of the great reliance upon the personal evaluative judgments of the individual staff members, what criteria were set up for participation in the program, and what running checks, if any, were made upon the efficiency of staff personnel?

The total impression of the OSS assessment program that one gets from this book is that it was instituted under an arbitrarily selected philosophy of assessment, conducted by individuals enthusiastically committed to this point of view, and sustained by faith rather than concrete evaluative data. This is not to say that it was not a good program, nor that a better one could have been developed. It is merely to wish that belief were reinforced by scientific data. The reader *probably* will agree when the authors say " . . . something definitely useful has come out of these expert labors. But it is not possible to say that they have added anything to our knowledge of the components and determinants of personality." He would be happier, however, had he been presented with more adequate data upon which to base his judgment.

WILLIAM A. HUNT.

Northwestern University.

TOMKINS, S. S. *The Thematic Apperception Test.* New York: Grune & Stratton, 1947. Pp. vii+291.

Clinical psychologists have for some time anticipated an authoritative, book-length discussion of the Thematic Apperception Test to bring order in the chaos of minor research and diverse treatment of thematic material following Murray's original presentation. Tomkins' book is the work of a philosopher and scholar, as well as an able clinician. It contributes an original approach to the problems of scoring, interpretation and research, based upon a rationale which considers the instrument in the larger setting of scientific methodology with special reference to the science of personality. It is presented as "a workbook, rather than as a compilation of established doctrine" and offers a system of interpretation which in many respects is essentially new. Perhaps a peculiarity of the TAT is the fact that no system so far presented has gained general use, and it seems unlikely that this contribution will

supplant eclecticism, or replace more flexible approaches of the type set forth by Rapaport. The scoring is based upon a set of variables ("vectors," "levels," "conditioners," and "qualifiers") which, in the hands of anyone but their originator, are likely to prove unduly labored.

Of particular interest from the standpoint of methodology are the opening chapters which undertake an analysis of the analytic process engaged in by the interpreter, for the purpose of demonstrating that "The interpretation of TAT stories may employ canons of inference long accepted by other sciences." This examination of the logic of deduction is a valuable corrective to our sometimes myopic research for substantiation of hypotheses in a complex matrix of validating criteria. Nevertheless, some students of personality are likely to find in Tomkins an over-emphasis upon reductionistic, cause-effect analysis, impressive in its grammarian thoroughness, but likely to obscure, especially for beginners, less abstract and more illuminating interdependencies in the data. The approach is from microscopic to macroscopic analysis, whereas the reverse is the method of preference for many.

The practicing clinician is likely to find greatest satisfaction in the later chapters in which, after the author has described his tools, he demonstrates their use. The sections on personality diagnosis, in which the various regions of family, love and sex, social relationships, work and vocational settings are discussed in significant dimensions, are rich in case material, with interpretation masterfully handled, and new diagnostic insights supplied.

In the writer's opinion, a highly important contribution is Tomkins' discussion of repression. Here he takes issue with orthodox psychoanalytic conceptions regarding the pathogenic nature of "deep" repression, enunciating a theory by which the seriousness of conflict is judged not by depth of repression, but by the intensity and extensity of conflicting wishes and the degree of deadlock between them. This concept has import for therapy and for theory of personality, as well as for the prognostic potentialities of the TAT.

The book opens with a useful summary of important research in the Thematic Apperception Test and closes with a discussion of the test as an adjunct to therapy. Although it is unlikely to fill the need for an integrative text in the field, it marks the coming of age of an important psychodiagnostic device, and by reason of its thorough exposition and systematic contribution deserves to become an essential reference in TAT research development.

HELEN SARGENT.

Northwestern University.

BOOKS AND MATERIALS RECEIVED

ANDREWS, T. G. (Ed.) *Methods of psychology.* New York: Wiley, 1948. Pp. xiv+696. $5.00.

BLUEMEL, C. S. *War politics and insanity.* Denver, Col.: World Press, 1948. Pp. 117.

BORING, E. G., LANGFELD, H. S., AND WELD, H. P. (Eds.) *Foundations of psychology.* New York: Wiley, 1948. Pp. xv+613. $4.00.

BURTT, H. E. *Applied psychology.* New York: Prentice-Hall, 1948. Pp. x+812. $7.35.

CHURCHMAN, C. W. *Theory of experimental inference.* New York: Macmillan, 1948. Pp. xi+287. $4.25.

COTTRELL, L. S., AND EBERHART, SYLVIA. *American opinion on world affairs—in the atomic age.* Princeton: Princeton Univ. Press, 1948. Pp. xxi+152. $2.50.

DAILEY, J. T. (Ed.) *Psychological research on flight engineer training.* Army Air Forces Aviation Psychology Program, Research Reports, Report No. 13. Washington: U. S. Govt. Printing Office, 1947. Pp. vii+220. $1.25.

DAVIS, R. A. *Educational psychology.* New York: McGraw-Hill, 1948. Pp. x+340. $3.00.

DEUTCH, ALBERT. (Ed.) *Sex habits of American men.* New York: Prentice-Hall, Inc., 1948. Pp. x+235. $3.00.

EVANS, R. M. *Introduction to color.* New York: Wiley, 1948. Pp. x+328. $6.00.

EYSENCK, H. J. *Dimensions of personality.* New York: Cambridge Univ. Press, Macmillan, 1948. Pp. x+308. $5.00.

FRANK, L. K. *Projective methods.* Springfield, Ill.: Thomas, 1948. Pp. vii+86. $2.75.

GALLUP, G. E. *A guide to public opinion polls.* (Rev. Ed.) Princeton: Princeton Univ. Press, 1948. Pp. xxiv+117. $2.50.

GEDDES, D. P., AND CURIE, ENID. *About the Kinsey report.* New York: New Amer. Library, 1948. Pp. 168. $0.25.

HEADLEE, RAYMOND, AND COREY, BONNIE W. *Psychiatry in nursing.* New York: Rinehart, 1948. Pp. xi+296. $3.50.

HINSHAW, DAVID. *Take up thy bed and walk.* New York: Putnam, 1948. Pp. xvi+258. $2.75.

HSU, F. L. K. *Under the ancestors' shadow.* New York: Columbia Univ. Press, 1948. Pp. xiv+317. $3.75.

JONES, ERNEST. *Shakespeare's "Hamlet," with a psychoanalytic study.* New York: Funk & Wagnalls, 1947. Pp. 180. 8s 6d.

KEYSERLING, G. H. *Das Buch vom Ursprung.* Buhl-Baden: Roland-Verlag, 1947. Pp. 371.

LEVI, A. W. *General education and social studies.* Washington: American Council on Education, 1948. Pp. xviii+336. $3.50.

MACKINNON, D. W., AND HENLE, MARY. *Experimental studies in psychodynamics. A laboratory manual and experimental materials.* Cambridge: Harvard Univ. Press, 1948. Pp. ix+177. $5.00.

METTLER, F. A. *Neuroanatomy.* (2nd Ed.) St. Louis: Mosby, 1948. Pp. 491.

MOORE, T. V. *The driving forces of human nature.* New York: Grune & Stratton, 1948. Pp. viii+456. $6.50.

PENNINGTON, L. A., AND BERG, I. A. *An introduction to clinical psychology.* New York: Ronald Press, 1948. Pp. xv+566. $5.00.

RUCH, F. L. *Psychology and life.* (3rd Ed.) Chicago: Scott Foresman, 1948. Pp. xvi+763.

SARTRE, J. P. *The psychology of imagination.* New York: Philosophical Library, 1948. Pp. 282. $3.75.

SCHWEITZER, ALBERT. *The psychiatric study of Jesus.* Boston: Beacon Press, 1948. Pp. 75. $2.00.

SHERIF, MUZAFER. *An outline of social psychology.* New York: Harper, 1948. Pp. xv+466. $4.00.

SPROTT, W. J. H. *General psychology.* New York: Longmans Green, 1948. Pp. ix+457. $3.25.

THORPE, L. P., AND KATZ, BARNEY. *The psychology of abnormal people.* Pp. xvi+877. $6.00.

WOLBERG, L. R. *Medical hypnosis.* (2 volumes.) New York: Grune & Stratton, 1948. (Vol. I: *The principles of hypnotherapy*, pp. xi+449, $5.50; Vol. II: *The practice of hypnotherapy*, pp. vii+513, $6.50.)

Guidance handbook for secondary schools. Division of Research and Guidance of the Office of Los Angeles County Superintendent of Schools. Los Angeles: California Test Bureau, 1948. Pp. xxi+243. $3.00.